Faculty of Science, Technology, Engineering and Mathematics
M140 Introducing statistics

Unit 5 Relationships

Contents

Introduction

In Units 2 and 3, we looked at prices and incomes and attempted to answer the question: *Are people getting better or worse off?* You have learned several statistical techniques for summarising a batch of data and for comparing two batches. In Unit 4, you saw how to choose a sample for a survey and you were introduced to ideas about random sampling.

In this unit, we are not going to attempt to answer any particular question, but we are going to investigate relationships between two variables. Scatterplots can be used to picture such relationships, and they were used for this purpose in Unit 1. For instance, in Subsection 2.1 of Unit 1 some data on the quantity of fertiliser and yield of wheat-grain produced were given in a table. A scatterplot was then used to explore the relationship between the quantity of fertiliser applied and the yield of grain. The scatterplot is reproduced in Figure 1. It shows that grain yield increased as more fertiliser was applied. Moreover, the plotted points lie roughly in a straight line, suggesting that a straight line could be used to model the relationship between the two variables. This happens quite often when two variables are related – an aim of this unit is to give a way of calculating a straight line that best represents the relationship between the variables.

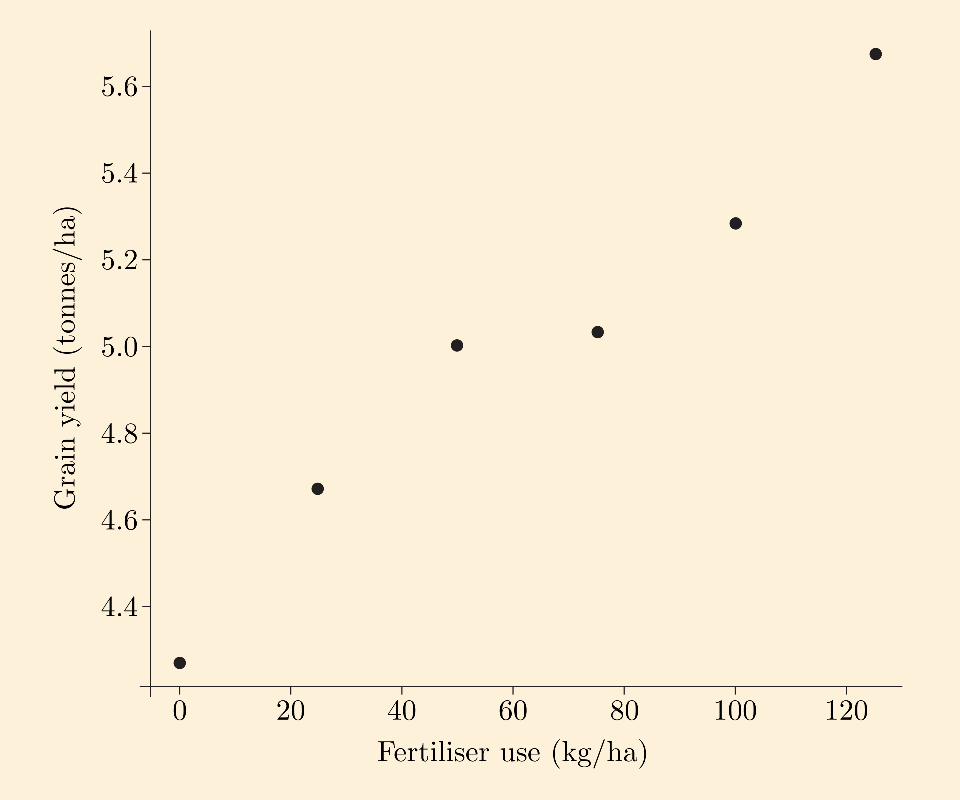


Figure 1 Grain yield by fertiliser use

We also explore other information that scatterplots yield both about relationships and about individual data points, paying particular attention to unusual points that are a long way from the main body of data. After drawing a straight line on a scatterplot, we address the question of how to tell when the line is a good way of representing the relationship between the two variables.

Section 1 describes what is meant by a *relationship* between two variables; it introduces scatterplots in detail, in particular addressing which variable should be plotted on which axis. In Section 2, scatterplots are used to characterise different types of relationship and identify unusual data points. Drawing an appropriate line on a scatterplot to represent a relationship is first considered in Section 3, where the line is drawn 'by eye'. Calculating the *least squares fit line* to model the relationship is the topic of Section 4. Uses to which this line can be put are described in Section 5.

Section 6 directs you to the Computer Book. You are also guided to the Computer Book at the end of Subsection 4.1. This work is critical to the unit and it is strongly recommended that you do it at this point in the text. However, you will be reminded in Section 6 in case you have not completed it by then.

1 Relationships and scatterplots

Statisticians refer to quantities like age, height or price, which vary from one individual or one purchase to another, as **variables**. Two variables are said to be **related** if knowing the value of one of the variables provides information about the value of the other variable. In this unit we shall be looking at various questions that can be asked about related variables.

- How can we investigate whether two variables are related?
- How can we describe a relationship between two variables numerically?
- What use can we make of a numerical description of a relationship?
- How can we interpret or explain a relationship?

First, though, we consider more closely what constitutes a relationship between two variables.

1.1 What is a relationship?

Suppose you are the membership secretary of a sports club for which the minimum age of entry is 10 years. You have to apply this rule, but obviously you do not want to upset potential members. One day two girls arrive and say they would like to join. One is about ten centimetres taller than the other. You ask the taller girl how old she is and she replies that she is 10. So you guess that the other girl will be too young to join; however, when asked, she says that she is 12. You are surprised because you based your guess on the fact that taller girls are usually older. In other words, there is a **relationship** between age and height of girls. It is not a perfect relationship because a 12-year-old girl may be shorter than a 10-year-old, and, assuming the two girls are telling the truth, this is what happened in the case above.

Suppose your son offered to do the weekly shopping at the supermarket and you asked him to get 10 kg of potatoes, although usually you only buy 5 kg of potatoes. You would not know exactly how much these would cost, because the price varies between varieties and from week to week. Also, a bag containing 10 kg of potatoes usually costs a little less than two 5 kg bags. However, you would probably have some idea of how much these would cost, as the weight of potatoes provides a guide to this.

These two situations both involve relationships between two variables. In the case of the sports club, knowing the girls' heights enabled you to guess at the girls' ages (wrongly, as it turned out). The relationship applies both ways; knowing a child's age would give you information about his or her height. It is not precise information. For example, if you were told that a girl was eight years old today, you could not say that she was exactly 1.25 metres tall. However, you could be fairly certain that she would be shorter than a 12-year-old girl, and you could say (given the appropriate information) that she would probably be between 1.18 and 1.32 metres tall.

In the potato example, you probably thought in terms of price per kilogram. However, price is just a way of describing the relationship between weight of potatoes and amount of money paid. Generally, the more you buy, the more you pay.

Activity 1 Investigating height and age

To start you thinking about what is involved in learning about a relationship, try answering some questions related to the sports club situation above.

(a) How would you investigate the relationship between height and age in children?

(b) How might you describe the numerical relationship between height measurements and age values of children?

(c) Would you expect to see the same sort of relationship between age and height in adults as you would see in children?

1.2 Linked data

Let us now turn to a different example of a relationship – that between car ownership and unemployment. That is, if we know the rate of unemployment in an area, does that tell us anything about the amount of car ownership in the area?

Example 1 Unemployment in Bedfordshire; car ownership in Merseyside

In the UK ten-yearly census, all households are required to complete a detailed return, and this provides information on many topics, including car ownership and unemployment. Each household records (amongst many other things) the number of cars owned by members of the household, the number of men aged 16–74 in the household, and the number of those men who are unemployed on the date of the census. A 10% sample of these data is analysed and the results are published in the form of percentages for every town and region in the UK. Tables 1 and 2 show some of the results from the 2001 census. Table 1 shows the percentage of men unemployed in four regions of Bedfordshire. Table 2 shows the percentage of households with no car in five regions of Merseyside.

Table 1 Male unemployment in Bedfordshire

Bedfordshire	Percentage of men unemployed
Bedford	3.99
Luton	4.82
Mid Bedfordshire	2.07
South Bedfordshire	2.73

(Source: HMSO (2004) *Census 2001: Key Statistics for Local Authorities in England and Wales*, Table KS09b)

Table 2 Access to cars in Merseyside

Merseyside	Percentages of households with no car
Knowsley	41.76
Liverpool	48.28
St Helens	30.48
Sefton	31.00
Wirral	30.34

(Source: HMSO (2004) *Census 2001: Key Statistics for Local Authorities in England and Wales*, Table KS17)

Activity 2 Considering unemployment and car ownership

Do the figures in Tables 1 and 2 above provide any information about a possible relationship between household car ownership and male unemployment rates?

As you saw in Activity 2, to investigate the relationship between car ownership and unemployment, we need linked data giving both percentages for a number of towns. For convenience, 'town' means town or small region for the remainder of this example.

Linked data

Data are said to be **linked** when two or more variables are recorded for the same sampling units.

When there are two variables, linked data are also often referred to as *paired data*.

Data from the UK Census in 2001 includes unemployment rates and rates of car ownership for towns in Great Britain. Because of the large number of towns in Great Britain, a stratified sampling scheme was used to select the data below. The sampling was limited to England and used the main regions as strata. One town or small region was selected from each of West Midlands, North West, Yorkshire and the Humber, North East, East Midlands, South West, and East, and three towns were selected from the South East region, which is the most populated. London and the major cities were omitted because they might not be typical of the country as a whole. Both percentages (the 'variables' in this case) were recorded for each of the ten towns (the 'sampling units'); the linked data are shown in Table 3.

Table 3 Male unemployment and car ownership for ten towns in England

Town	% males unemployed	% households with no car
Alnwick, North East	4.59	21.6
Vale Royal, North West	3.55	17.2
Rotherham, Yorkshire and the Humber	5.19	29.7
Rutland, East Midlands	1.75	13.6
Dudley, West Midlands	5.27	25.3
Norwich, East	5.61	35.5
Bracknell Forest, South East	2.25	14.5
Rother, South East	3.00	20.8
Mole Valley, South East	1.84	13.1
West Dorset, South West	2.14	16.9

(Source: HMSO (2004) *Census 2001: Key Statistics for Local Authorities in England and Wales*, Tables KS09b and KS17)

In Table 3 each row gives the percentage of males unemployed for that town and the percentage of households with no car for the *same* town. So you can get some idea of the relationship between male unemployment and car ownership by just looking at the two columns of numbers. For example, Rutland and Mole Valley have the lowest percentages in both columns, whereas Rotherham and Norwich both have a high percentage of male unemployment and households with no car.

Note that ‘relationship’ is not meant to imply anything about *causality*. That is, a relationship between male unemployment and lack of cars does not imply that unemployment causes lack of car ownership, or that a lack of cars causes unemployment.

Activity 3 Linked or not linked?

For each of the two sets of data described below, state whether the data are linked data or not.

(a) Measurements of heights of two groups of children: one group of twenty year-6 children and one group of twenty year-7 children.

(b) Measurements of height for one group of twenty year-7 children, both one year ago and now.

1.3 Scatterplots

In the previous subsection you saw that linked data can be displayed in a table. Relationships in the data can be explored by looking in the table for patterns. However, a scatterplot gives us a better impression of linked data, making it easier to spot patterns. Figure 2 shows a scatterplot of the unemployment and car ownership data given in Table 3.

Alternative expressions for scatterplot are **scattergram**, **scattergraph** and **scatter diagram**.

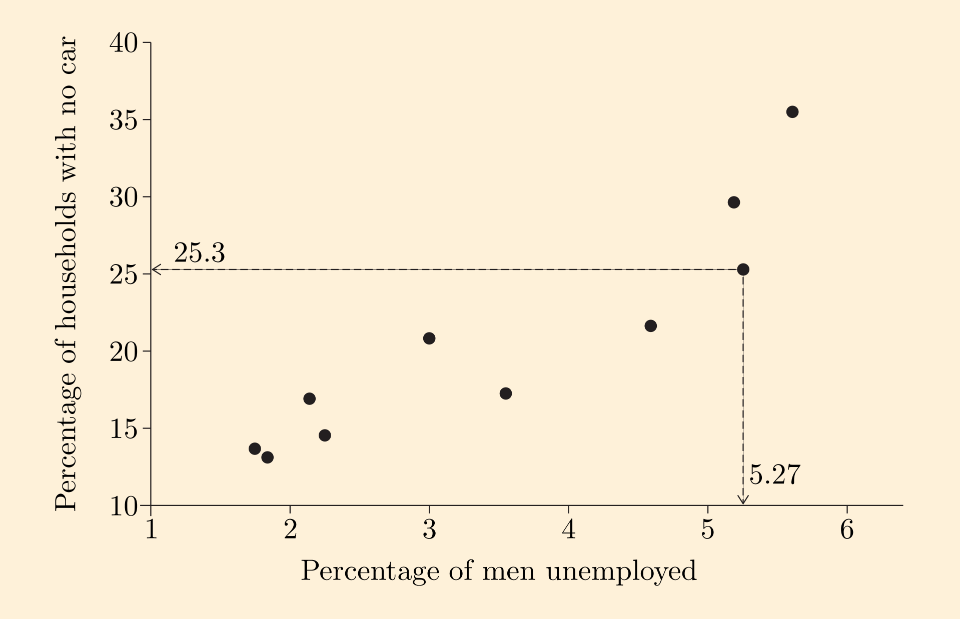


Figure 2 A scatterplot of car ownership against unemployment

In Figure 2, the horizontal axis (the x-axis) is labelled ‘Percentage of men unemployed’, and the vertical axis (the y-axis) is labelled ‘Percentage of households with no car’. These axes represent the two columns of data we are displaying in the plot. The scales have been chosen to cover the range of the data (from 1.75 to 5.61 for the percentage of males unemployed, and from 13.1 to 35.5 for the percentage of households with no car). The numbering of scales is done with numbers that can be written down using just a few significant figures. As happens in many scatterplots, the plotted scales do not start at zero. Instead the range of values plotted for each axis is chosen so that points on the scatterplot cover as much of the area of the scatterplot as possible.

Computer software for drawing scatterplots is usually able to choose the scales automatically.

There are ten points marked on the scatterplot. These points represent the ten towns in Table 3. The position of each town’s point is given by the two values in the corresponding row on Table 3. For example, in Dudley at the time of the census in 2001, 5.27% of men were unemployed and 25.3% of households had no car. So the point representing Dudley is placed at the position corresponding to 5.27 along the horizontal axis and 25.3 along the vertical axis. The position of points on a scatterplot can be written concisely using **coordinates**. For example, the values 5.27 and 25.3 are the coordinates of the point representing Dudley: 5.27 is called the **first coordinate** or **x-coordinate**, and 25.3 is called the **second coordinate** or **y-coordinate**.

A wide variety of symbols can be used to mark points on a scatterplot, including dots or crosses.

In order to emphasise that the values 5.27 and 25.3 are the coordinates for a point on the scatterplot, it is common to write the two values side by side, separated by a comma and enclosed in brackets like this:

$$(5.27, 25.3).$$

The first number in the bracket is the value along the horizontal axis and the second number is the value along the vertical axis. It is important always to write the numbers in this order in the brackets. The coordinates $(25.3, 5.27)$ would tell us that for some town in 2001, 25.3% of men were unemployed and that 5.27% of households had no car.

One way of remembering which way round to write the numbers in brackets is that in the alphabet ‘h’ comes before ‘v’, and so the coordinate along the *h*orizontal axis comes before the coordinate on the *v*ertical axis.

Activity 4 Exploring a scatterplot

(a) Using the data in Table 3 write down the coordinates for the following two towns: Vale Royal and Rother.

(b) Which town is represented by the point in the top rightmost corner of Figure 2?

(c) Describe in words what the scatterplot tells you about the relationship in this batch of data.

Can you see the upper points of my scatterplot?

Activity 4 asked to you describe the scatterplot in Figure 2. The description and interpretation of scatterplots will be considered in more detail in Section 2.

1.4 Response and explanatory variables

One important aspect of constructing a scatterplot has not yet been mentioned: which variable to put on the x-axis and which variable to put on the y-axis. When investigating the relationship between two variables it often happens that the values taken by one variable can be partly explained using the values taken by the other variable. For example, the height of a child can be partly explained by the child's age. In such situations, the variable being explained (such as child's height) is called the *response* variable, and the variable doing the explaining (such as child's age) is the *explanatory* variable.

Sometimes it makes more sense to think that the value of one variable depends, at least in part, on the value of the other variable. In this case the response variable is the variable that depends on ('responds to') the other variable, and the variable on which the response variable depends is the explanatory variable. To illustrate, suppose you are given some linked data on petrol consumption of a car: various speeds and the miles per gallon when the car travelled at each speed. In this case, miles per gallon is the response variable, as it depends, to a certain extent, on the speed of the car, which is the explanatory variable.

Finally, sometimes the value of one variable is to be predicted, and the value it takes is partly related to the value of the second variable. For example, you might want to predict the miles per gallon you will obtain if you drive at 65 mph. Then the variable to be predicted is the response variable and the second variable is the explanatory variable.

Explanatory and response variables

A **response variable** is the variable that is being explained or whose value depends on other variables. It is also the variable to be predicted if predictions are to be made. It is sometimes known as the *dependent* variable.

An **explanatory variable** is the variable that is doing the explaining or is the variable on which the response variable depends. It is sometimes known as the *independent* variable.

By convention, on a scatterplot the explanatory variable is put on the x-axis and the response variable is put on the y-axis.

Example 2 Blood pressure

Suppose that you are given some linked data on blood pressure that consists of blood pressure measurements on patients before and after a treatment. In this case, the blood pressure measurement after treatment is the response variable, as it depends to a certain extent on the blood pressure before treatment, which is the explanatory variable. It does not make sense to think that someone's blood pressure before the treatment can be changed by changing their blood pressure after the treatment.

So, on a scatterplot the blood pressure before the treatment would be plotted along the x-axis, and the blood pressure after the treatment would be plotted along the y-axis.

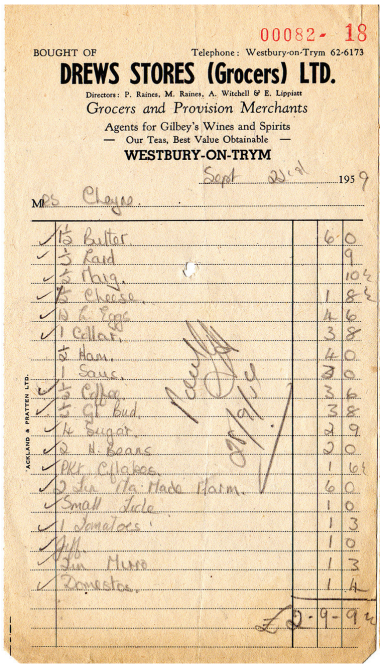
A sphygmomanometer, an instrument used to measure blood pressure

Example 3 Household expenditure

Suppose in a survey of household expenditure, 12 households are asked to record their total expenditure for one week and also to note what items they bought.

If a household has a low income, it spends this on necessities, including food, and cannot afford luxuries. However, when income increases, more money will be spent on luxuries. Although the household will probably spend more on food, maybe paying for higher quality, the increase is proportionately less, and so the percentage of total income spent on food falls.

So when we draw a scatterplot of total expenditure and percentage of total income spent on food, total expenditure is the explanatory variable and percentage spent on food is the response variable. This means that total expenditure would be put on the x-axis, and the percentage of total expenditure spent on food would be put on the y-axis. This makes sense as a household is very unlikely to decide what percentage of its total expenditure should go on food before working out what its total expenditure should be.



00082- 18

BOUGHT OF Telephone: Westbury-on-Trym 62-6173

DREWS STORES (Grocers) LTD.

Directors: P. Raines, M. Raines, A. Witchell & E. Lippiatt

Grocers and Provision Merchants

Agents for Gilbey's Wines and Spirits

— Our Teas, Best Value Obtainable —

WESTBURY-ON-TRYM

Sept 22nd 1959

Mrs Cheyne.

✓ 1½ Butter		6	0
✓ ½ Lard			9
✓ ½ Marg.			10½
✓ ½ Cheese.		1	8½
✓ ½ L Eggs		4	6
✓ 1 Cellar.		3	8
½ Ham.		4	0
1 Saus.		3	0
✓ ½ Coffee.		3	6
✓ ½ Gt Bud.		3	8
✓ 4 Sugar.		2	9
✓ 2 H. Beans		2	0
✓ Pkt Cflakes.		1	6½
✓ 2 Lin Ma Made Marm.		6	0
✓ Small Tide		1	0
✓ 1 Tomatoes		1	3
✓ Jiff.		1	0
✓ Tin Mono		1	3
✓ Domestos.		1	4
	£2 ·	9 -	9½

28/9/59

ACKLAND & PRATTEN LTD.

Some old household expenses

In most experimental situations, it is usually clear which is the explanatory variable. An experiment often consists of choosing values of an explanatory variable (for example, amount of fertiliser applied, dose of a drug given to patients, temperature of an industrial process) and then observing the effect on the response variable (for example, yield of tomatoes, blood pressure of patients, strength of a manufactured component).

Sometimes, though, the use that will be made of the data determines which variable is the response variable and which is the explanatory variable. If we wish to forecast one variable when the other takes a particular value, the variable we wish to *forecast* is regarded as the response. For example, if a married man's height is to be predicted from the height of his wife, then the man's height is the response and the wife's height is the explanatory variable. These roles are reversed if a married woman's height is to be predicted from the height of her husband. If the use of the data is unspecified, then it is arbitrary which of their heights should be plotted on the x-axis and which on the y-axis.

Activity 5 Which variable would you plot on the x-axis?

For each of the following cases, if you were asked to draw a scatterplot of the data, which variable would you choose as the x-coordinate? Give a reason for your choice.

(a) In order to investigate the effects of different amounts of fertiliser on the yield of tomatoes, ten tomato plants of the same variety were each given a different amount of the same fertiliser. The data consist of the amount of fertiliser and the weight of tomatoes for each plant.

(b) The data consist of the numbers shown in Table 3 (Subsection 1.2): the percentage of males unemployed and the percentage of households with no car in a random sample of ten towns in England.

(c) For a series of water companies in the UK, the average consumption by households with water meters and the average consumption by households without water meters were collected.

Some tomatoes growing on a plant

Exercises on Section 1

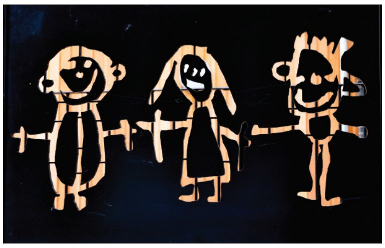
Exercise 1 Linked or not?

For each of the following pairs of variables, state whether you think they are linked or not. Justify your opinion.

(a) The heights of a group of twenty 5-year-old children in one school and the weights of twenty 5-year old children in a different school.

(b) The heights of twenty 5-year old children and weights of another twenty 5-year-old children, all from the same school.

(c) The heights and weights of twenty 5-year-old children.



A depiction of some children

Exercise 2 Identifying explanatory and response variables

For the following pairs of linked variables, discuss which variable could be regarded as the response variable and which as the explanatory variable.

(a) Average house price and calendar year.

(b) Average hourly wage earned by men and average hourly wage earned by women, in different sectors of the economy.

(c) In a study to predict employment rates, the unemployment rate in different countries and the employment rate in those countries.

2 Interpreting scatterplots

In Section 1, the use of linked data to investigate relationships in data was introduced. You also saw that such data can be displayed graphically as a scatterplot. In this section, we shall investigate what can be learned from looking at a scatterplot.

When interpreting a scatterplot, we are only concerned with a general overall relationship. That is, the general pattern set by the vast majority, if not all, of the points. Any points that do not fit with the general pattern might be treated separately. We shall return to this point in Subsection 2.4.

2.1 Positive and negative relationships

Look again at the scatterplot given in Figure 2 (Subsection 1.3), which shows the relationship between percentage of men unemployed and percentage of households with no car.

The points on the scatterplot do not lie exactly on a straight line. This means that if we were told the percentage of unemployed men in a town, we would not know the exact percentage of households without a car. However, knowing the percentage of unemployed men does tell us something about the percentage of households without a car. As was noted in Activity 4 (Subsection 1.3), there is a tendency for towns with a low unemployment rate to also have a low percentage of households with no car. Similarly there is a tendency for towns with a high unemployment rate to have a high percentage of households with no car.

This is more clearly seen by looking at the shaded area shown in Figure 3. The shaded area is chosen so that it contains all the points.

By concentrating on the shaded area instead of the individual points, the general pattern between the percentage of men unemployed in a town and the percentage of households without a car becomes clearer. The area slopes upwards from left to right, so towns with a low unemployment rate, like Mole Valley, have a low percentage of households with no car, while towns that have a high unemployment rate, like Rotherham, also have a high percentage of households with no car.

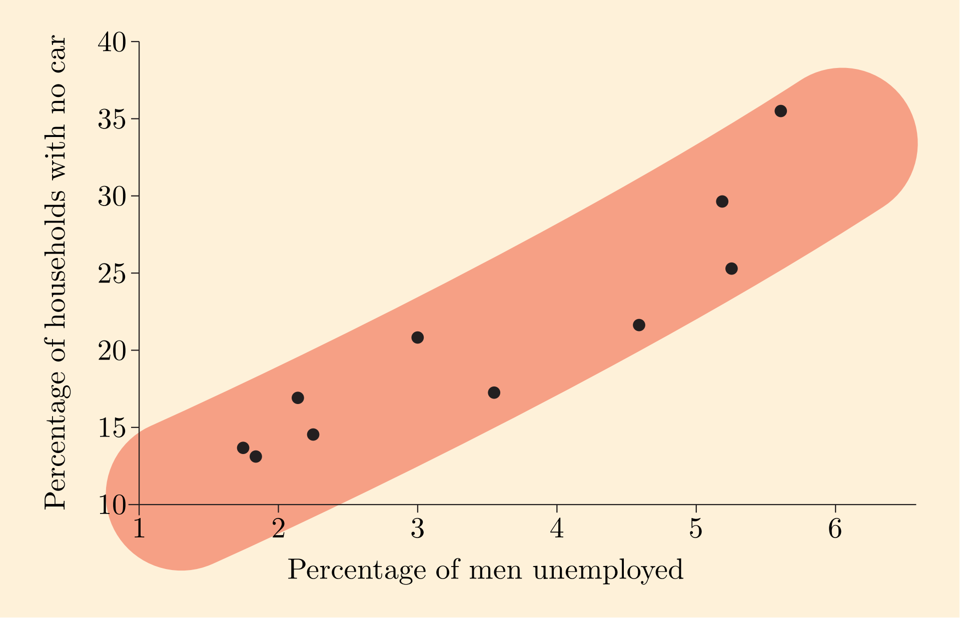


Figure 3 Percentage of males unemployed and percentage of households with no car in ten towns

Lots more positive relationships

When the area enclosing the points slopes upwards from left to right, as in Figure 3, then we say that the variables are **positively related**. So, Figure 3 implies that the male unemployment rate and the percentage of households without a car are positively related.

Figure 4 shows the scatterplot of some data relating to weekly household expenditure for 12 regions and nations in the UK. The data points are again enclosed in a shaded area.

This time the area slopes downwards from left to right. High weekly expenditure is associated with a low percentage of expenditure on food and non-alcoholic drink, and low weekly expenditure is associated with a high percentage of expenditure on food and non-alcoholic drink. When large values of x are usually associated with small values of y, and small values of x are associated with large values of y, as in Figure 4, the variables are said to be **negatively related**.

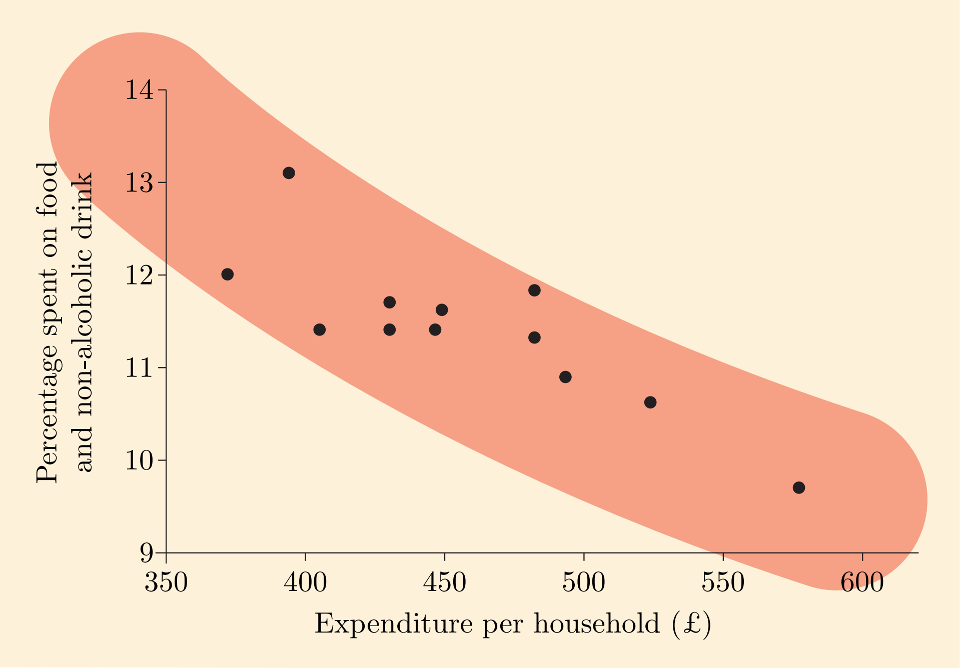


Figure 4 Expenditure per household and percentage spent on food and non-alcoholic drink

Positive and negative relationships

On a scatterplot, variables are said to be **positively related** if low values of x are associated with low values of y, and high values of x are associated with high values of y.

That is, if points tend to slope *upwards* from left to right, then the variables are *positively* related.

Variables are said to be **negatively related** if low values of x are associated with high values of y, and high values of x are associated with low values of y.

That is, if points tend to slope *downwards* from left to right, then the variables are *negatively* related.

Activity 6 Positive or negative relationship?

In Figures 5 and 6 below, are the variables positively or negatively related?

(a) A dataset of 50 observations.

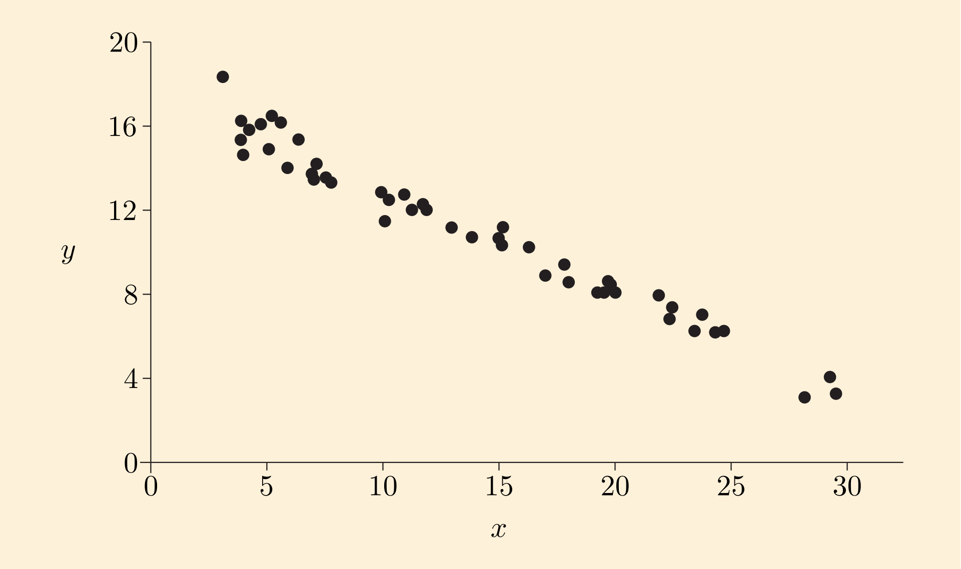


Figure 5 A scatterplot of some data

(b) A scatterplot of data from an experiment in kinesiology. A subject performed a standard exercise task at a gradually increasing level. The x-coordinate measures the amount of oxygen uptake, and the y-coordinate is the expired ventilation, which is related to the rate of exchange of gases in the lungs.

Note that the units for the oxygen uptake and expired ventilation would normally be included in the labelling of the scatterplot axes. However, in this case, the units are not known. The precise choice of units does not make a difference as to whether the relationship is positive or negative.

An example of the type of equipment that can be used to measure oxygen uptake

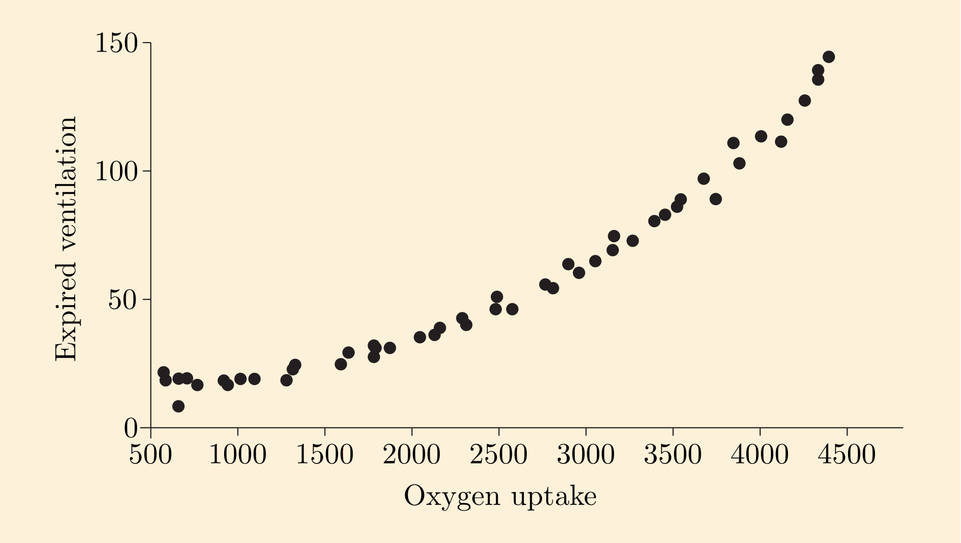


Figure 6 Data from an experiment in kinesiology

Not all pairs of variables are positively or negatively related. Figure 7 shows the number of adult road-user casualties in Scotland on weekdays, averaged over 2005 to 2009, for each hour of the day.

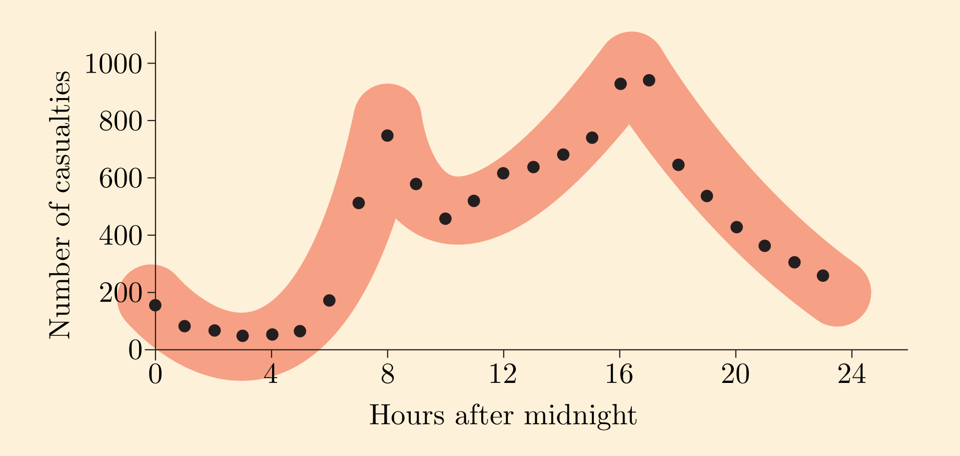


Figure 7 Casualties to adult road users on weekdays

You can see that there is a very pronounced pattern; the number is highest between 4 pm and 6 pm (16:00 and 18:00) when many adults are going home. The number is also quite high between 8 am and 9 am when many adults are travelling to work or doing the school run. On the other hand, the number is very low between 12 am and 7 am.

So there is a definite relationship between time of day and number of casualties, but it is much more complex than one that can be described simply as either positive or negative.

2.2 Linear and non-linear relationships

In the previous subsection, you saw that relationships on scatterplots can be described as positive, negative or neither. This subsection concentrates on another aspect of a relationship that can be investigated by looking at a scatterplot: whether it is linear or non-linear.

Example 4 Daily electricity usage and cost

In 2011, a household recorded the cost of their electricity usage on ten different days. The number of kilowatt hours (kWh) they used on each of these days along with the amounts they were charged are shown in Figure 8.

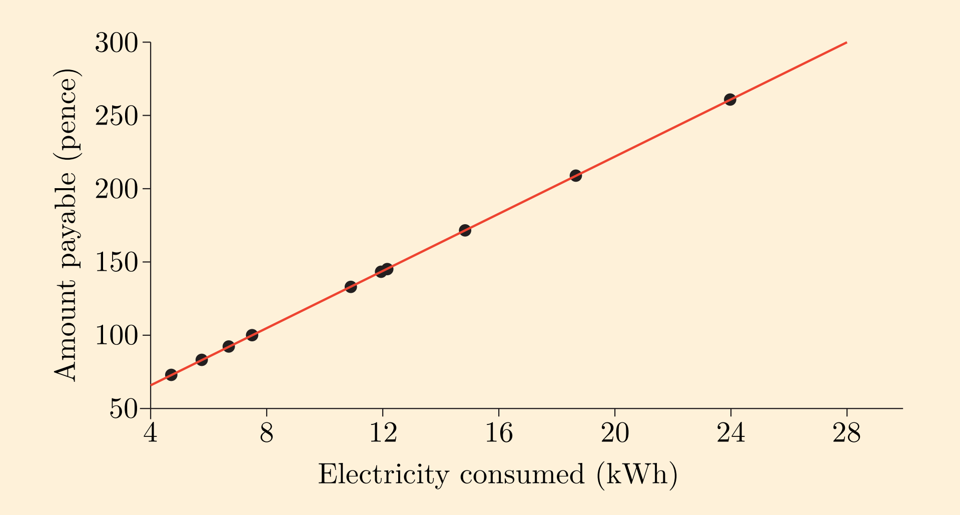


Figure 8 Daily cost of electricity for a household

Looking at the scatterplot, it is clear that there is a very precise relationship between the quantity of electricity consumed and the amount payable on the electricity bill: the data points lie on a straight line.

Many domestic electricity tariffs in the UK are made up of a fixed standing charge per day plus a charge per kilowatt hour used. For this household, the standing charge was 27 pence per day, and the cost per kWh was 9.7p. On day 1, this household's electricity cost $11.96 \times 9.7\text{p} = 116\text{p}$ (rounded to the nearest penny), so the total amount payable for that day was $27\text{p} + 116\text{p} = 143\text{p}$. If a household uses x kWh of electricity in a day, the total payable for that day is $27 + 9.7x$ pence. So if y is the amount payable in pence, then

$$y = 27 + 9.7x.$$

This is the equation of the straight line shown in Figure 8.

The relationship between electricity usage and amount payable in Figure 8 is represented by a straight line. So there is said to be a **linear relationship** between the electricity used and the amount payable.

The points need not lie exactly on a straight line for a relationship to be linear. Consider again the relationship between male unemployment and the percentage of households with no car. The shaded area of the scatterplot in Figure 3 (Subsection 2.1), and hence the underlying relationship, can be summarised by drawing a line through the middle of it as shown in Figure 9. This line happens to be a straight line, so this relationship is also said to be linear. More precisely, it is said to be a **positive linear** relationship, as the line goes up from left to right.

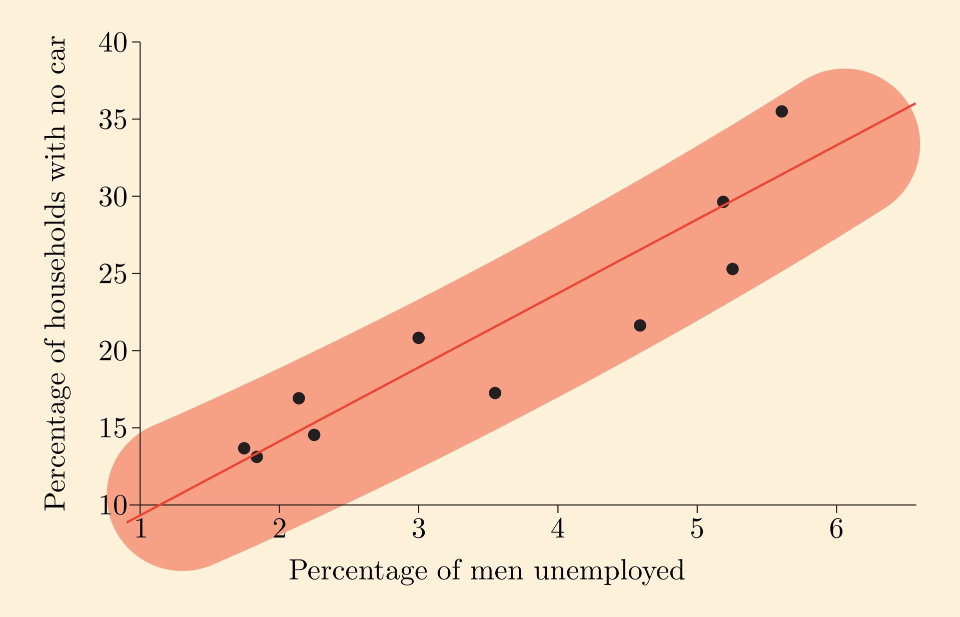


Figure 9 Percentages of males unemployed and households with no car in ten towns, with a line in the middle of the shaded area

It is also possible to draw a straight line going through the shaded area on the scatterplot of household expenditure given in Figure 4 (Subsection 2.1). However, this time the line would go down from left to right. So the relationship between household expenditure and the percentage spent on food is a **negative linear** relationship.

Now consider again the data from the experiment in kinesiology introduced in Activity 6 (Subsection 2.1). In Figure 10 a line has been drawn through the middle of the shaded area covering all the points.

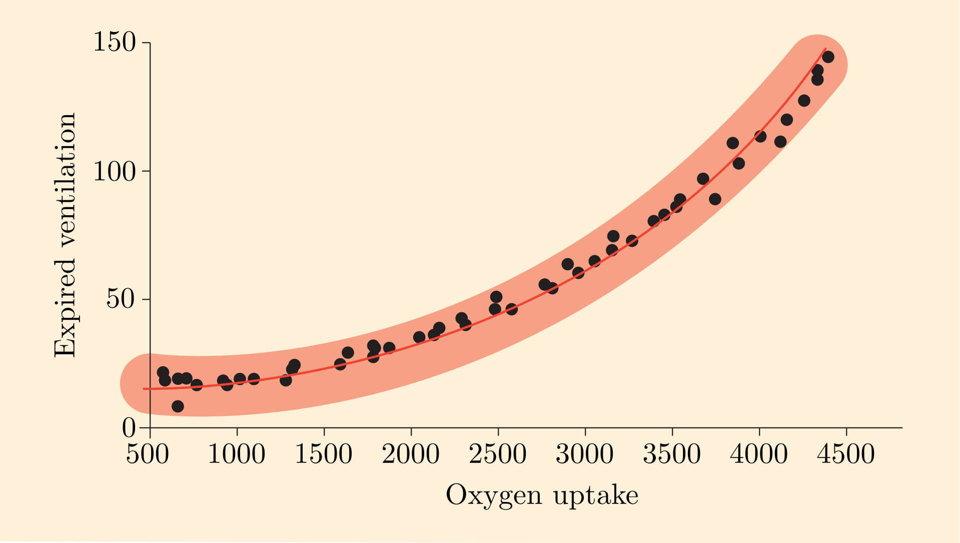


Figure 10 Data from an experiment in kinesiology with a line in the middle of the shaded area

Notice that the line in Figure 10 is curved, not straight. This is because the area containing the points is distinctly curved. So we say that the relationship between oxygen uptake and expired ventilation is **non-linear**.

Linear and non-linear relationships

A relationship is said to be linear if it can be summarised reasonably well by a straight line.

A relationship is said to be non-linear if it can be summarised reasonably well by a curve but not by a straight line.

Activity 7 Linear or non-linear?

Figure 11 is a scatterplot of data from an ultrasonic calibration study. Is the relationship between the variables linear or non-linear?

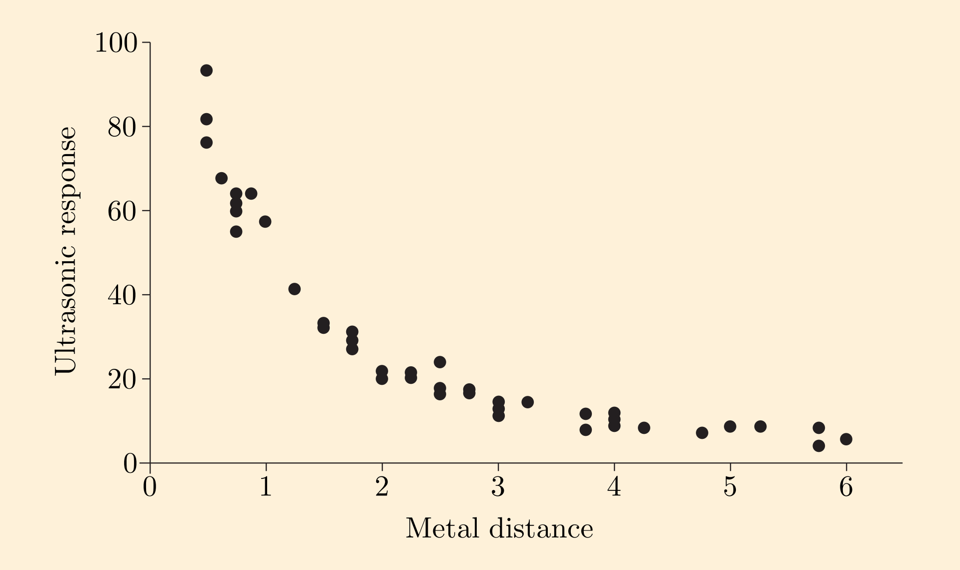


Figure 11 Data from an ultrasonic calibration study

2.3 Strong and weak relationships

Sometimes the general pattern formed by points on a scatterplot is very clear: if a line were drawn on the plot summarising the general pattern, then all the points would lie close to the line. In such cases we say that there is a **strong relationship** between the two variables. On the other hand, the general pattern might be difficult to pick out. Then the relationship between the two variables is said to be a **weak relationship**.

Example 5 A strong relationship

Look again at the data from the kinesiology experiment introduced in Activity 6 (Subsection 2.1). This scatterplot is reproduced in Figure 12 for convenience.

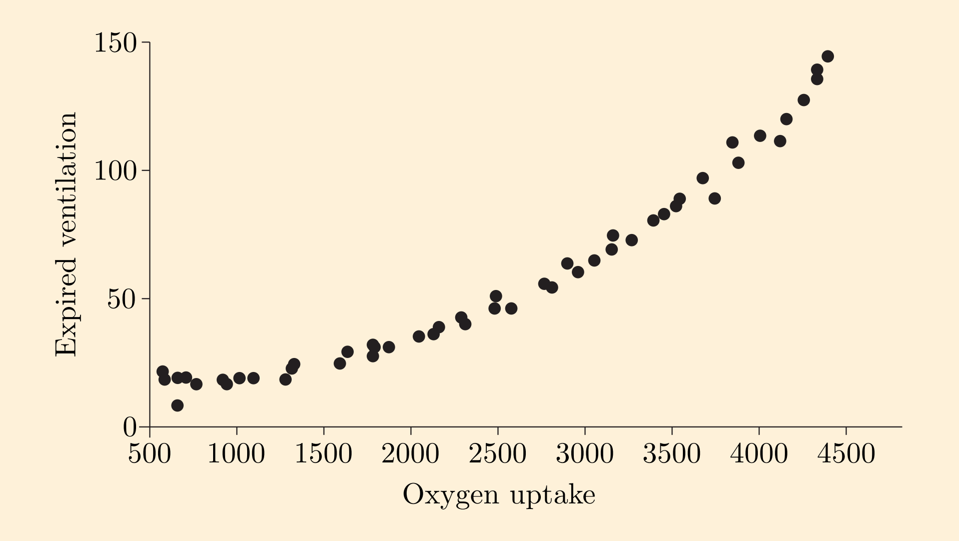


Figure 12 Data from an experiment in kinesiology

Notice that the points form a clear general pattern, curving upwards as you move from left to right. All the points lie close to this general pattern. Thus there is a strong relationship between oxygen uptake and expired ventilation.

Example 6 A weak relationship

Data were collected on the water consumption of customers from 22 water companies in the UK in 2008/09. The data are plotted in Figure 13.

Notice that in this scatterplot the overall pattern is not very clear. The points go generally up as you move from left to right. However, there is lots of scatter around whatever trend there is. So there is a weak relationship between the average consumption in metered households and the average consumption in unmetered households served by the same company.

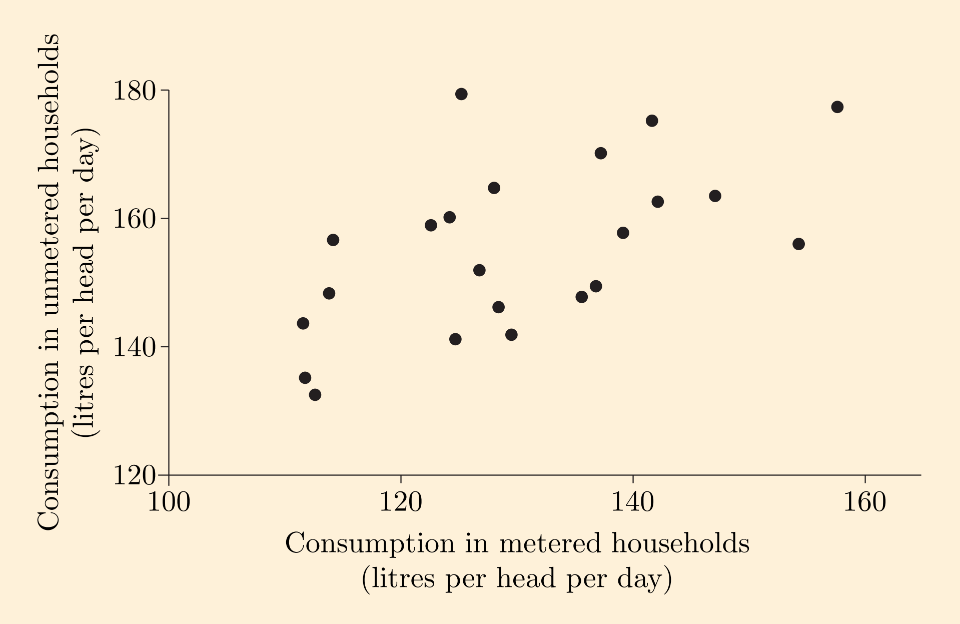


Figure 13 Average water consumption in metered and unmetered households

Strong and weak relationships

A relationship is said to be **strong** when all the points on a scatterplot lie close to a line.

A relationship is said to be **weak** when the points only loosely follow a line.

Deciding when a relationship is clear enough to be a ‘strong’ relationship is a subjective judgement. Sometimes the best that can be done is just to say whether a relationship on one scatterplot looks stronger than the relationship on another plot.

Activity 8 Comparing strength of relationships

Order the scatterplots in Figures 14, 15 and 16 according to the strength of the relationship between the variables, from strong to weak.

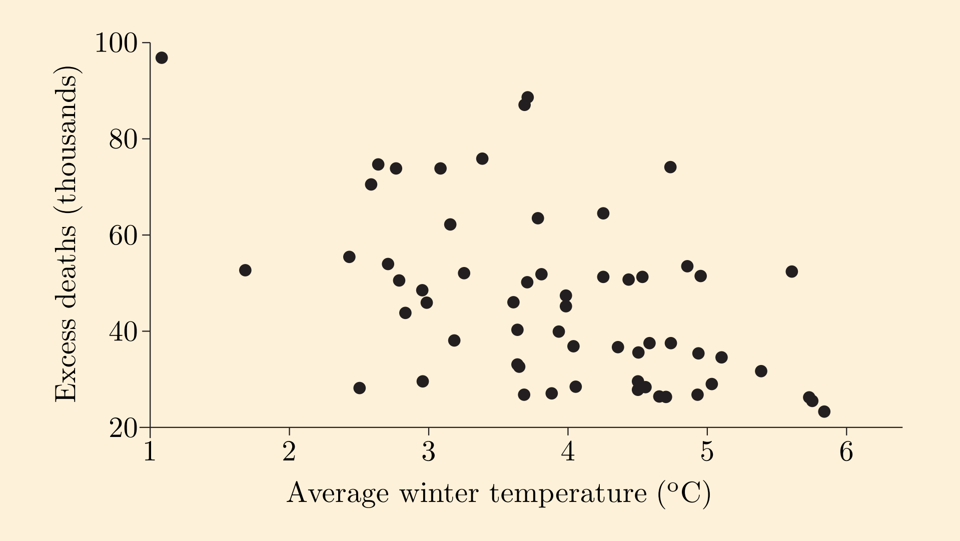


Figure 14 Estimated number of ‘excess’ deaths in winter (over and above the average for the rest of the year), plotted against average winter temperature, for Great Britain for each year from 1952 to 2010.

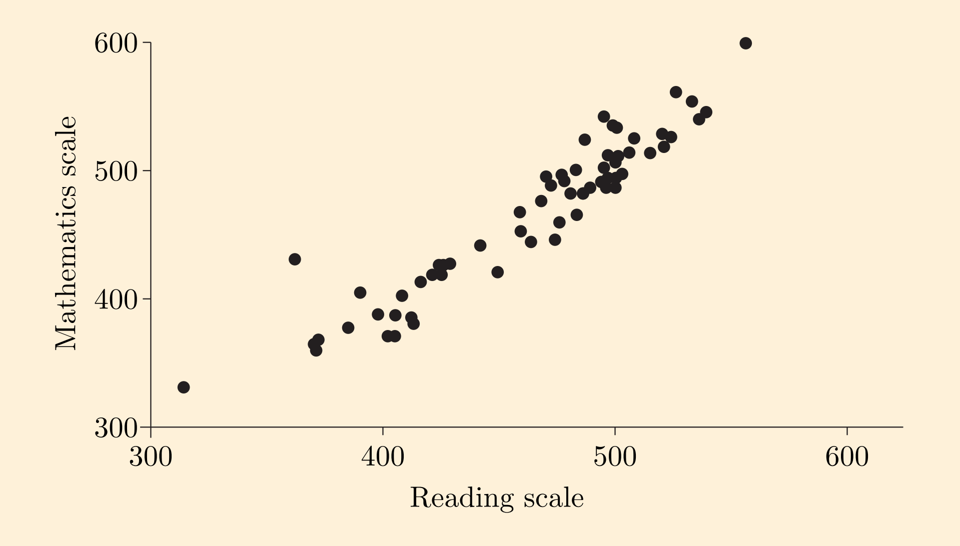


Figure 15 Average performance on a mathematics scale plotted against average performance on a reading scale, for 15-year-old students from different countries in 2009.

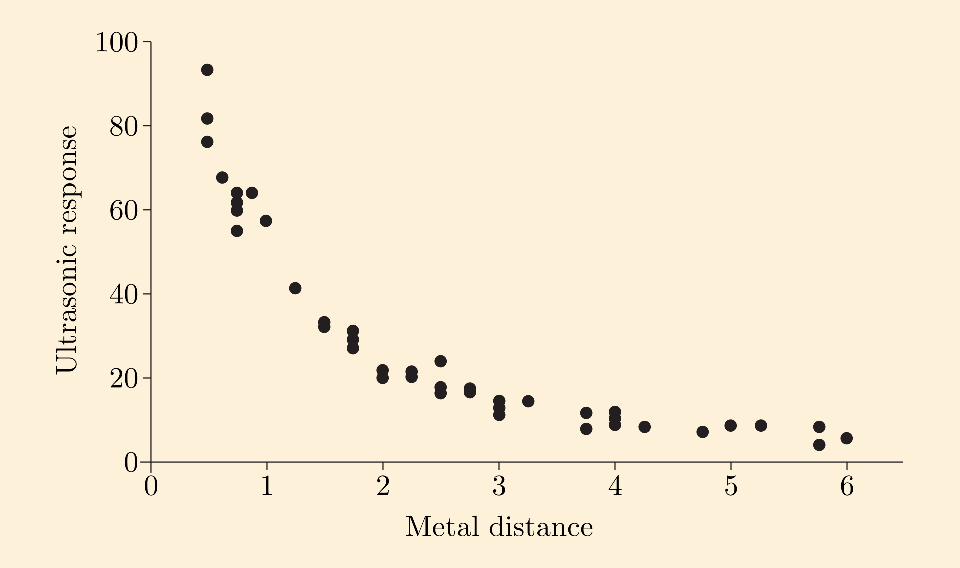


Figure 16 Data from the ultrasonic calibration study introduced in Activity 7 (Subsection 2.2)

Activity 9 Describing a relationship

Look at the data displayed in Figure 17. This scatterplot displays further information about the ten towns listed in Table 3 (Subsection 1.2) and considered in Figure 9 (Subsection 2.2). The x-axis corresponds to the percentage of employed residents working in the manufacturing industry, and the y-axis corresponds to the percentage of households living in owner-occupied houses.

What relationship between the variables can you observe in this scatterplot?

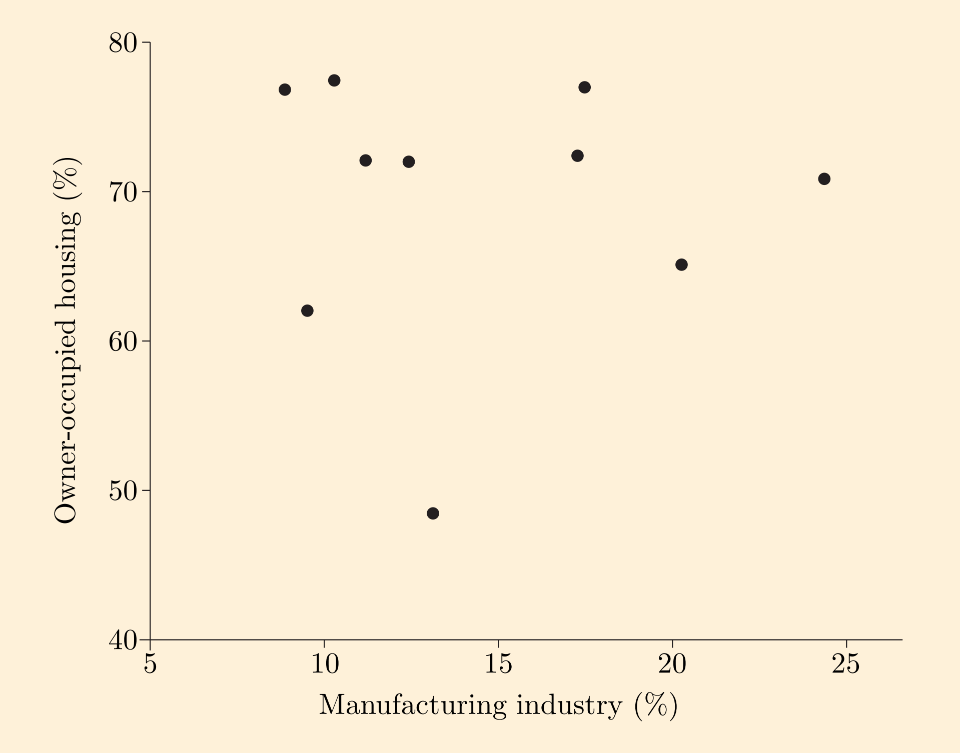


Figure 17 Percentage of employed residents working in manufacturing and percentage of households in owner-occupied houses

The scatterplot given in Activity 9 is an example of no relationship between two variables. If you were told, for example, that 14.4% of the workforce of Milton Keynes were employed in manufacturing industries, this would not help you at all in estimating the proportion of households who own their own homes. (Actually, it was 65.2% in 2001.)

There is said to be **no relationship** between two variables when knowledge of the value of the explanatory variable does not provide information about the value of the response variable.

2.4 Unusual points

In this subsection, there are some final comments about interpreting scatterplots and they relate to unusual points: sometimes one or two data points do not appear to follow the same pattern as the rest of the points.

For example, look at Figure 18, which shows a scatterplot of the percentage of the population aged under 16 in different regions of the UK in 2010 and the population densities (in people per km^2).

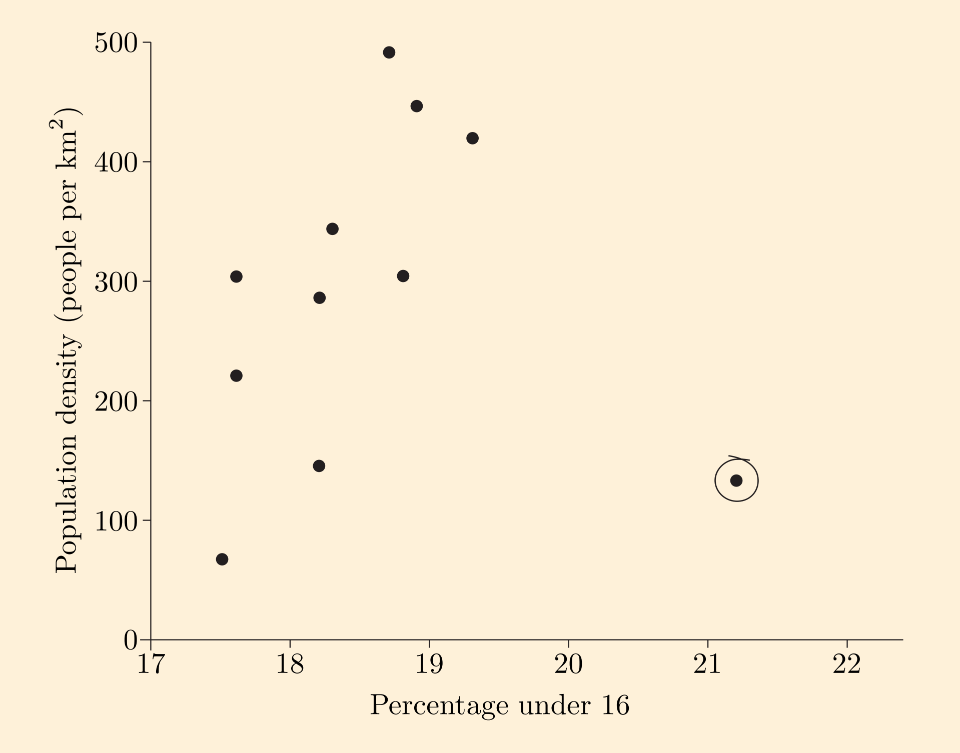
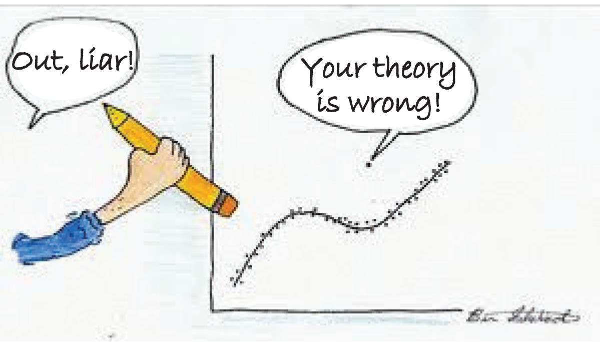


Figure 18 Percentage of the population aged under 16 and the population density

You can see that the ringed point, which represents Northern Ireland, does not follow the general pattern of the other ten regions. This point is called an *outlier*, because it is inconsistent with the main body of the data. This extends the definition of *outliers* given in Subsection 4.2 of Unit 1, where we considered only one variable at a time. The particular reasons for the investigation would determine whether or not Northern Ireland should be included when summarising the relationship between the two variables.

The x-value for Northern Ireland is unusual as it is much larger than those of all the other points. More generally, a point can be inconsistent with the main body of data even though neither its x-value nor its y-value is unusual – the *combination* of its x- and y-values can still place it a long way from other points and make it an outlier.



Finally, it should be noted that sometimes a point is only an outlier because a mistake has been made in the observation and/or recording of a data point. For example, if in the data used for Figure 18, the percentage of the population aged under 16 should really have been 18.2% instead of 21.2%, the point representing Northern Ireland would no longer appear to be an outlier. So looking for outliers on a scatterplot can help in data cleaning by highlighting parts of the data that are worth checking again for errors.

Data cleaning was introduced in Subsection 3.1 of Unit 1.

Activity 10 Spot the outlier

For each of the scatterplots below, how many outliers can you identify?

(a) The scatterplot of the average performance of 15-year-olds in different countries in 2009, introduced in Activity 8 (Subsection 2.3).

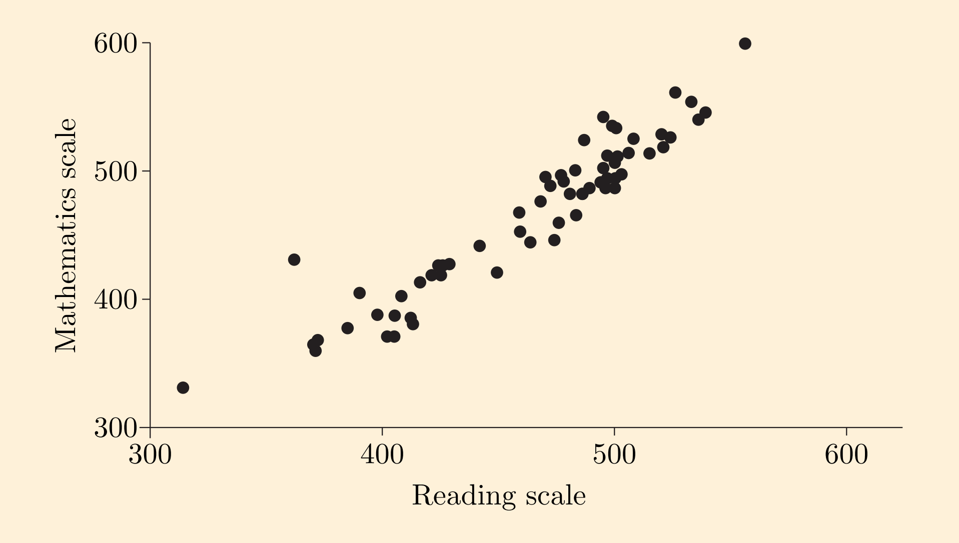


Figure 19 Student performance in reading and mathematics

(b) The scatterplot of some data relating to weekly household expenditure for 12 regions and nations in the UK, introduced in Figure 4 (Subsection 2.1).

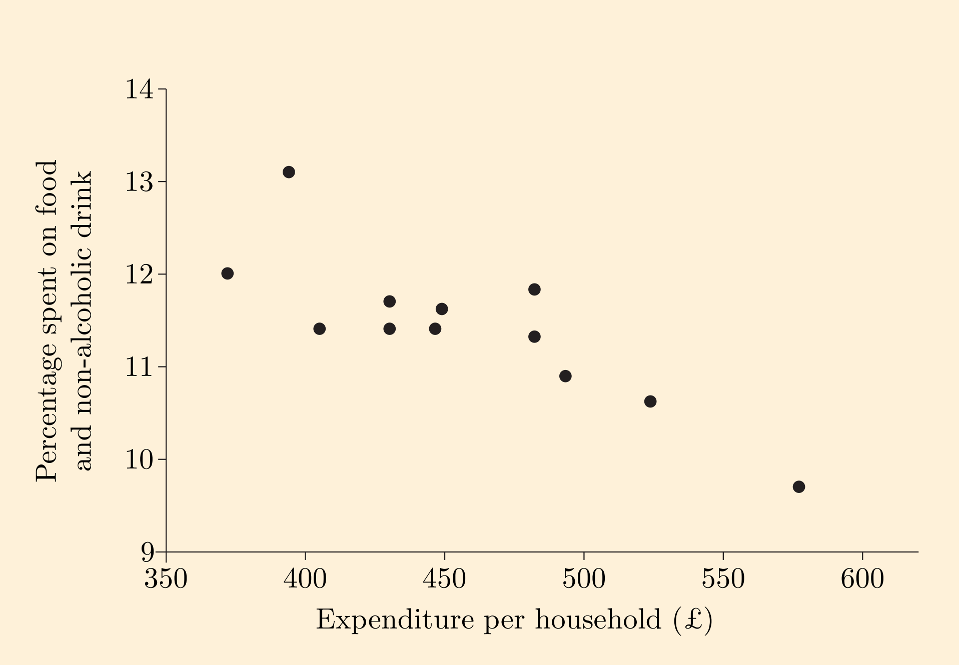


Figure 20 Percentage spent on food by households

In this section you have been learning to interpret scatterplots. Here is a checklist of things to consider.

Checklist for interpreting scatterplots

- Is the relationship positive, negative or neither?
- Is the relationship linear or non-linear?
- Is the relationship strong or weak?
- Are there any outliers?

You have now covered the material related to Screencast 1 for Unit 5 (see the M140 website).

Exercises on Section 2

Exercise 3 Average wages of men and women

Figure 21 is a scatterplot of men's and women's average hourly wages in different sectors of the UK economy. Interpret this scatterplot.

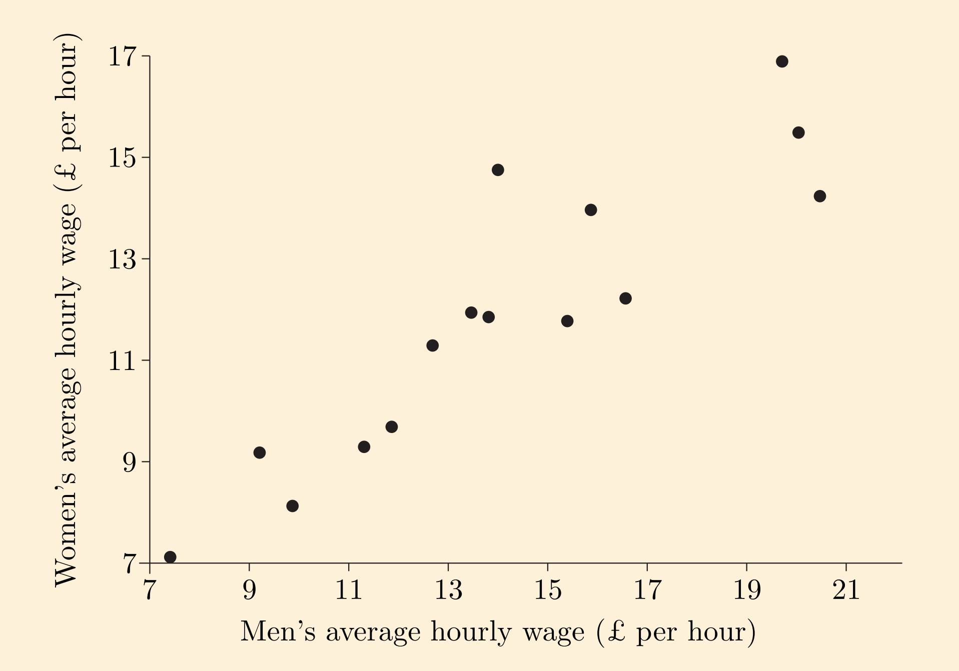


Figure 21 Average hourly wage for men and women in 15 sectors of the UK economy

Exercise 4 Investigating house prices over time

In Figure 22, a scatterplot of the average house price in the UK over the period 1991 to 2008 is shown. Using this scatterplot, comment on the pattern of house prices over this period.

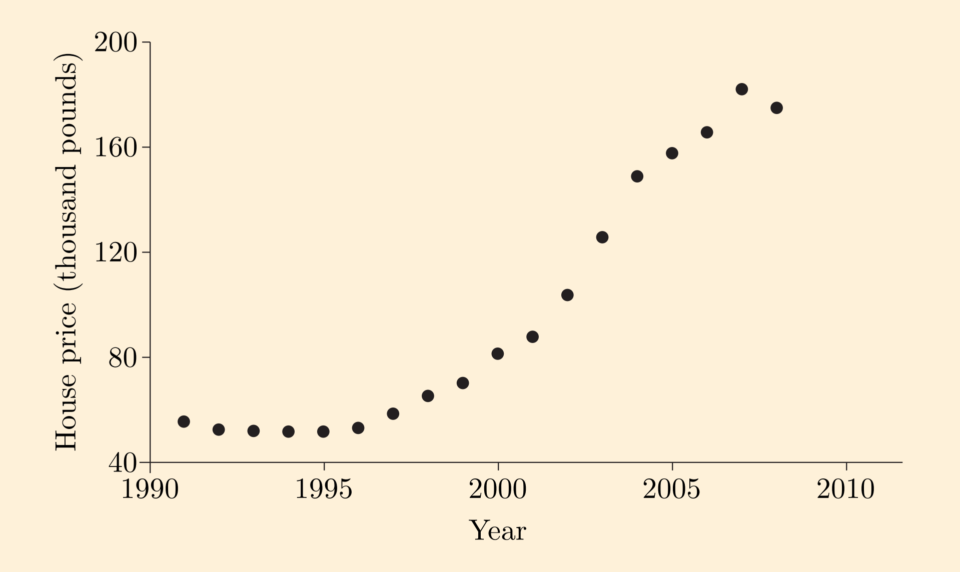


Figure 22 Average house price in the UK between 1991 and 2008

3 Scatterplots and lines

In Section 1, you were introduced to the idea of a relationship between two variables and learned how to present a relationship as a scatterplot. In Section 2, you saw how relationships may be positive, negative or neither, and that they can be linear or non-linear. In this section and Section 4, you will learn how to describe a relationship by adding a line and by calculating an equation. The line has many uses. For example, it will allow us to say something about the rate at which the response variable changes as the explanatory variable changes, and also to make informed predictions about the response when the value of the explanatory variable is known.

3.1 Drawing lines

You saw in Subsection 2.1 that many scatterplots could be approximately summarised by outlining an area on the scatterplot in which all, or nearly all, of the points lie. Often this area is long and narrow; for many datasets it is roughly straight, though it may be curved, as in Figure 6. However, sketching an area gives only a vague summary, and it would be useful to have a more precise measure.

In Example 4 (Subsection 2.2), we saw that the data points for daily electricity costs all lay exactly on a straight line. The line $y = 27 + 9.7x$ provides a precise summary of the data, and this line represents the relationship. In an **exact relationship** between two variables all the points lie exactly on a line which is either straight or follows a simple curve. Can we also represent an inexact relationship using a line? Well, often we can, just as we can represent the location of a batch of data by the median or the mean. Usually, hardly any of the data points are exactly equal to the mean, and, in the same way, a line used to represent a relationship will pass through very few, if any, of the data points. The purpose of the line is to represent the pattern that we can see in the points.

In statistics, the process of finding a line that best represents a relationship is known as **regression**.

Example 7 Summarising unemployment and car ownership

Data on percentages of males unemployed and households with no car were introduced in Subsection 1.2. In Subsection 2.2, you saw that the relationship between these quantities is approximately linear because the data can be summarised reasonably well by a straight line. One such line is shown in Figure 23. (Ways of choosing such a line will be considered later in this subsection and in Section 4.)

The line highlights the fact that towns with high male unemployment tend to have a relatively high percentage of households with no car, while those with low male unemployment tend to have a relatively low percentage of households with no car. The equation of the line shown in the figure is $y = 5.8 + 4.2x$.

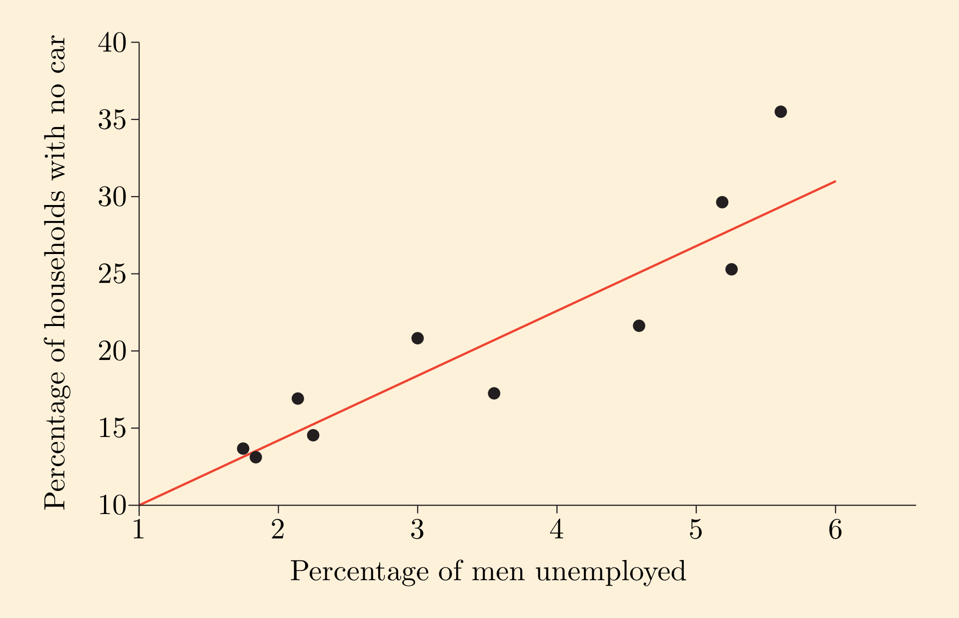


Figure 23 Percentage of males unemployed and percentage of households with no car, with straight line

Example 8 Summarising oxygen uptake

Recall from Subsections 2.1 and 2.2 that the scatterplot of oxygen uptake suggests a positive non-linear relationship between oxygen uptake and expired ventilation. This suggests that the data can be summarised by a curve that goes up as you move from left to right. One such curve is shown in Figure 24.

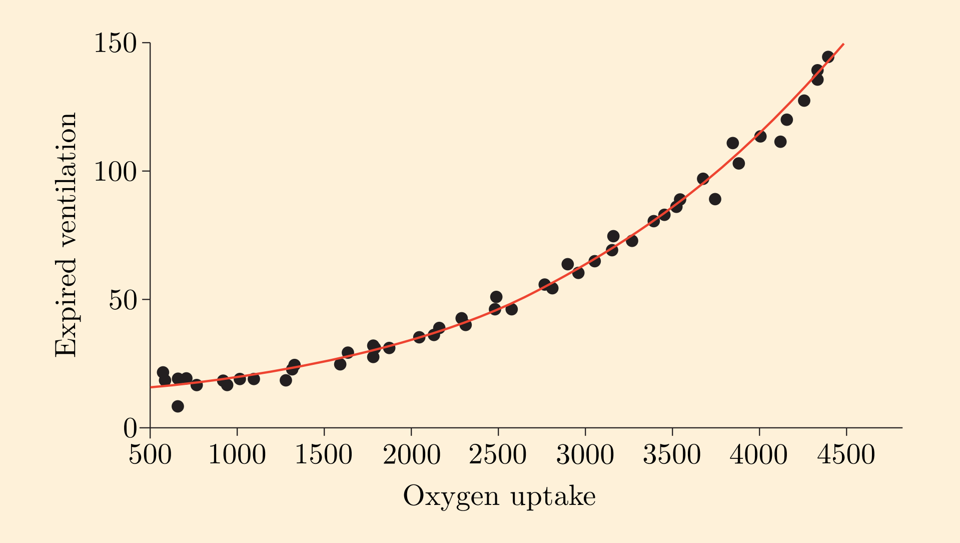


Figure 24 Data from an experiment in kinesiology, with curved line

Notice that the points are generally close to the curve. This is to be expected, as the relationship between oxygen uptake and expired ventilation is a strong one (as noted in Example 5, Subsection 2.3).

The lines in Examples 7 and 8 were both drawn by looking at the scatterplot and choosing a line that appears to provide a sensible model for the pattern made by the points.

Activity 11 Summarising data with a line

Suppose that the fictional data depicted in Figure 25 come from an experiment. In this experiment, an industrial process was run seven times, each time at a different temperature, and the yield recorded.

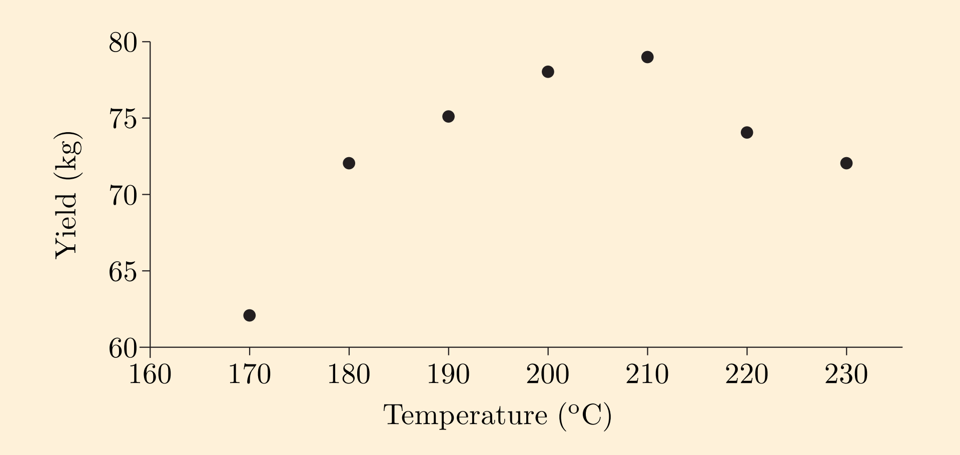


Figure 25 Yield from an industrial process

(a) Which variable is being treated as the response variable, and which as the explanatory variable? Does this choice seem reasonable?

(b) Briefly describe the relationship between temperature and yield shown in the scatterplot.

(c) Using your judgement, draw a line on Figure 25 which you feel provides a good summary of the data.

Your attempt at a line was probably slightly different from the one given in the solution to Activity 11(c). There is no single right answer to this question. However, your curve was probably of the same general inverted U-shape. The data certainly suggest that the yield is greatest when the temperature is round about 200–$210\,^\circ$C, but less if the temperature is either cooler or hotter. The line could have been drawn as a wiggly curve that went through all seven points on the scatterplot as in Figure 26.

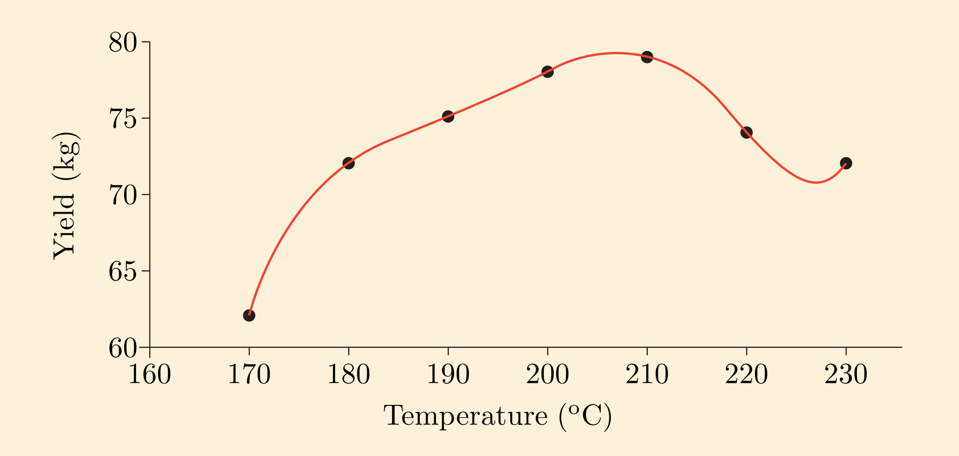


Figure 26 Yield from an industrial process, with a summary line going through all data points

However, it is important to bear in mind that the data were the result of one experiment. If the experiment had been repeated at $170\,^\circ$C, the yield might have been 61 or 64, say. The yield is subject to some experimental uncertainty, so a curve passing exactly through the points, as shown, is not the best way of summarising the data. A simpler curve is more useful.

To describe a scatterplot, we shall look for a simple curve that summarises the relationship. A straight line is the simplest sort of curve there is, so that will be used when it seems appropriate. For many of the scatterplots that we have looked at so far in this unit, a straight line does provide a reasonable summary of the pattern in the data.

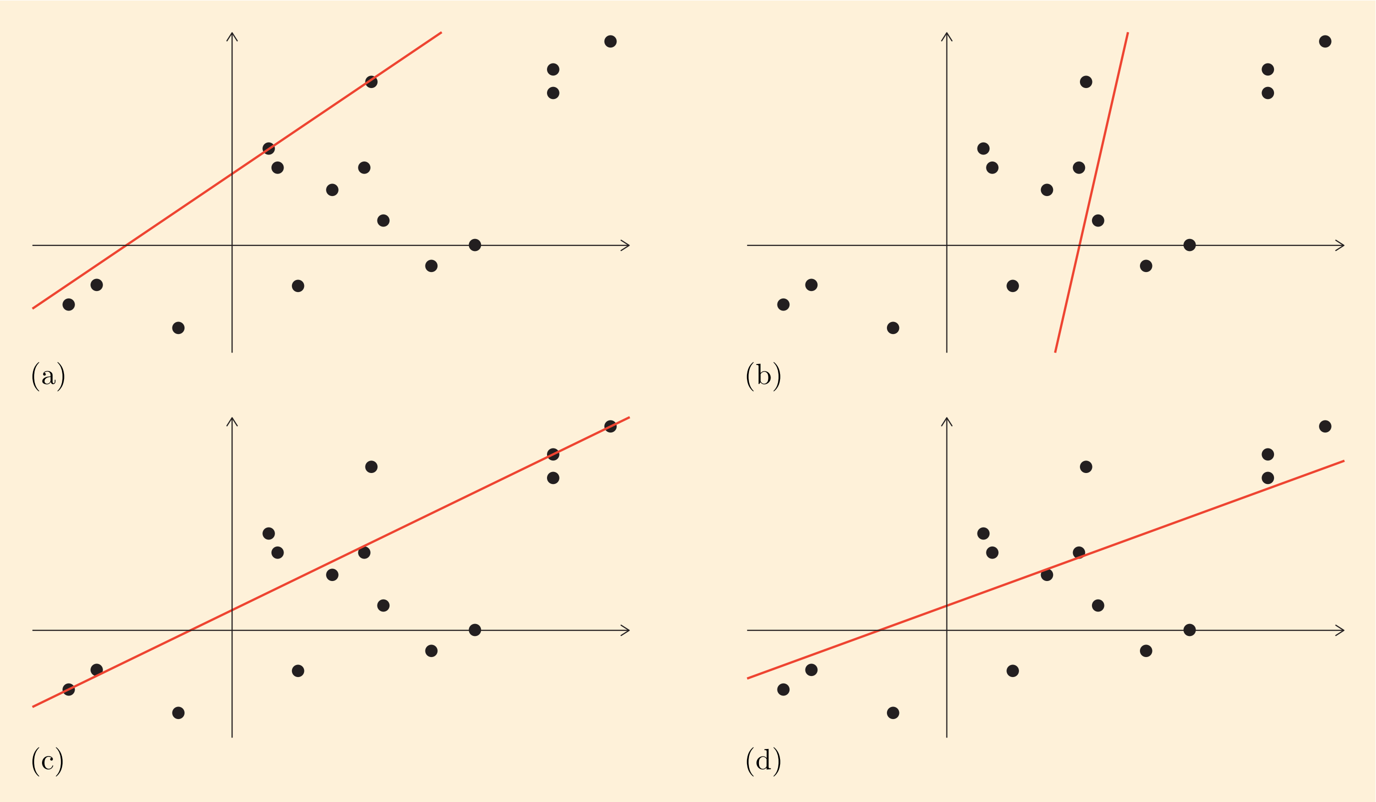
Describing a scatterplot

When summarising data on a scatterplot, the simplest adequate curve should be chosen. In many cases this amounts to choosing an appropriate straight line. This line is called the 'fitted line' or 'fit line'.

The process of choosing a straight line to draw is often called **fitting a line** to the data. There are many different ways to do this. One method, which you have already met, is simply to draw in the line that appears to give a good representation of the pattern in the data. This method, known as **fitting by eye**, can be perfectly adequate, particularly if the relationship is strong and all points are close to the line. But choosing a line that looks 'about right' is less easy when the relationship is weak and the points are widely scattered.

Activity 12 Comparing lines fitted by eye

Figure 12 shows four attempts at fitting a line by eye to the same set of data. In each case, say whether you think the line is a good choice or a poor choice. If you think it is poor, suggest how it might be improved.



You saw in the solution to Activity 12 that sometimes it is easy to say that a straight line is not a good fit. In Figure 12(a), the line was above all but two of the data points, and it would obviously be better if it had been lower and had passed through the middle of the cluster of points. Neither is the line in Figure 12(b) a good fit. Again, it does not follow the general pattern of the points, and it would be better if it were rotated clockwise to bring it near to the position of the line in either Figure 12(c) or (d). To choose a line that fits the general pattern of the data, it is sometimes helpful to use a transparent ruler and move it around until it appears to be in a good position.

However, it was hard to decide whether Figure 12(c) or (d) was a better choice. To get any further with the problem of deciding which straight line to draw, and whether a straight line does provide an adequate summary of the data, we need a more definite idea of what we mean by a *good* summary. We can think of this as whether the line provides a good fit to the data. This idea will be developed in the next subsections.

Fitting lines is good for you: the logo of a German purveyor of vitamin and mineral supplements!

3.2 Residuals

The basic idea for residuals uses an equation that will reappear later in this module: the **DFR equation**.

The DFR equation

This equation splits an observed ‘Data’ value into two parts: the ‘Fit’ and the ‘Residual’. These are linked in the following way.

$$\text{Data} = \text{Fit} + \text{Residual}.$$

The equation can be rearranged as

$$\text{Residual} = \text{Data} - \text{Fit}.$$

In other words, a residual is defined as the difference between a data value and a fit value.

Now suppose that when there are linked data, the 'Data' is taken to be the response variable. That is, for every point on the scatterplot, the 'Data' is the position of that point up the y-axis. And suppose that the 'Fit' is taken to be the position of a fitted line along the y-axis. That is, for every point on the scatterplot, the 'Fit' is the vertical position of the line, for the value of the explanatory variable. Then the 'Residual' is a measure of how far away each data point is from the fitted line. This is illustrated in Figure 27.

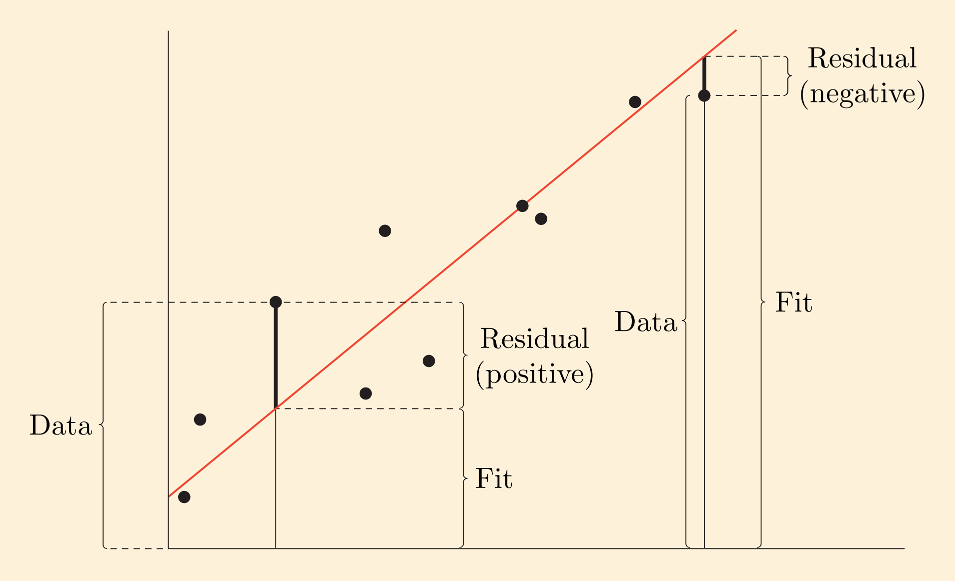


Figure 27 A scatterplot showing two residuals

Note that a 'Residual' is the *vertical* distance between a data point and the line. The 'Residual' is positive if the line is below the data point, and the 'Residual' is negative if the line is above the data point. If the 'Residual' is zero, the data point lies exactly on the fitted line. A 'Residual' close to zero indicates that the point is close to the fitted line, and a 'Residual' further away from zero indicates that the point is a long way from the line.

The reason for focusing on vertical distances is clearer when values of the explanatory variable are fixed by the experimenter. In the introduction to this unit, an experiment from Subsection 2.1 of Unit 1 was mentioned, in which fertiliser was applied and the subsequent yields of grain were recorded. The levels of fertiliser used in the experiment corresponded to at the levels of 0, 25, 50, 75, 100 and 125 kg/ha. In Figure 28, a vertical dashed line is plotted through each of these levels of fertiliser. Even while the grain is growing, we know a scatterplot will put a value for yield on each of these vertical lines.

In Figure 29, the data points have been added, together with the fitted line. The yields we should have expected while the grain was growing are the points where the vertical lines cross the fitted line – these points are the 'Fit' values. The short, thicker lines from the fitted line to the data points are the residuals.

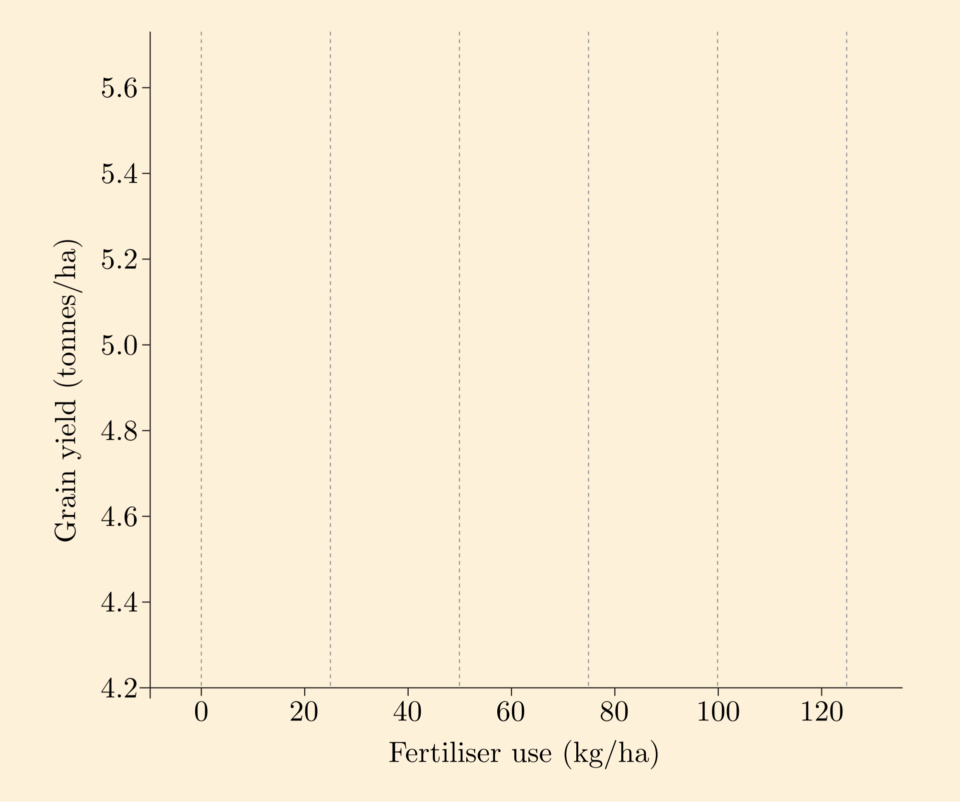


Figure 28 The levels of fertiliser that were applied

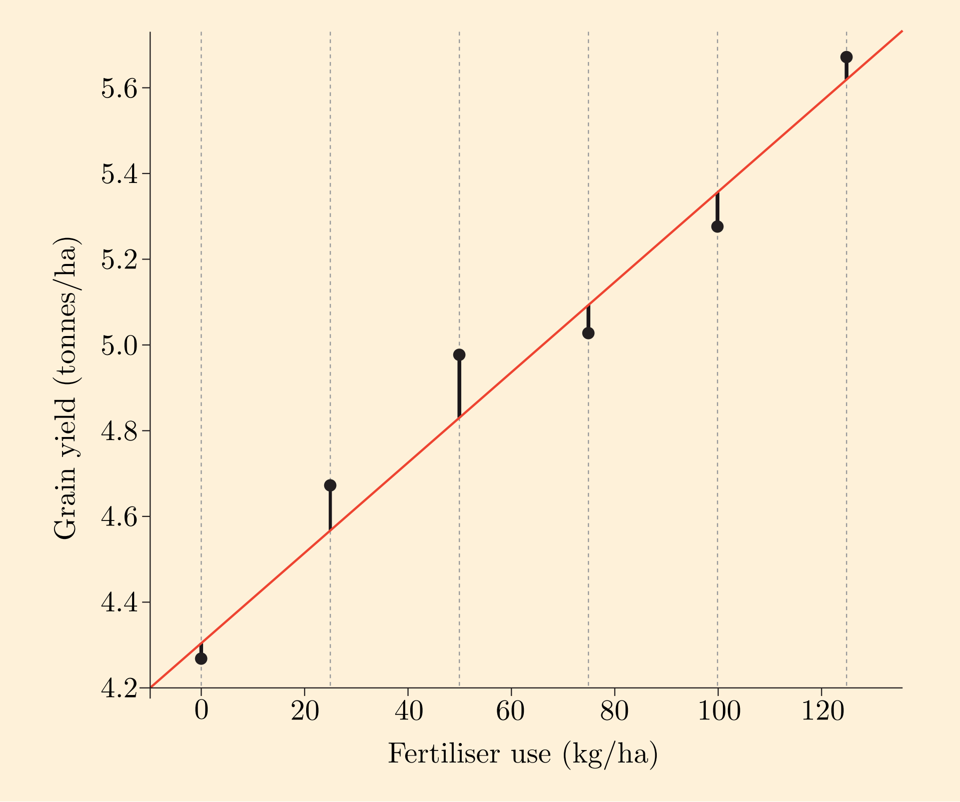


Figure 29 Residuals for the grain data

Example 9 Calculating residuals

Figure 30 is a scatterplot of linked data on the average expenditure per week and the average percentage spent on food and non-alcoholic drink, for 12 households. A line fitted by eye is also shown.

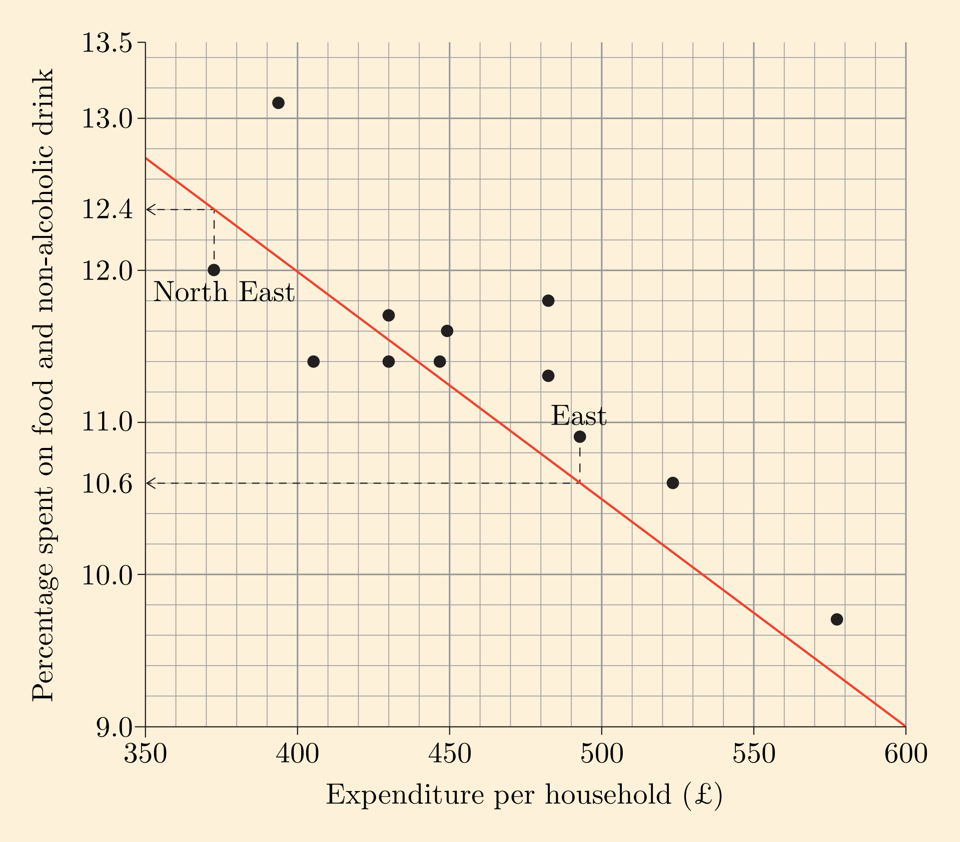


Figure 30 Average weekly household expenditure with a line fitted by eye

Consider the point in the figure that is labelled ‘North East’. This point has coordinates $(372.7, 12.0)$. That is the average expenditure per household is $372.70, and the average percentage spent on food and non-alcoholic drink is 12.0%. As the average expenditure per household is the explanatory variable, we are interested in the average percentage spent on food and non-alcoholic drink, y, when $x = 372.7$. In Figure 30, we can see that the point on the line which corresponds to $x = 372.7$ is $(372.7, 12.4)$. Thus 12.4 is the ‘Fit’ value. Using the DFR equation and applying it to the North East, we find

$$\begin{aligned}\text{Residual} &= \text{Data} - \text{Fit}\\ &= 12.0 - 12.4\\ &= -0.4.\end{aligned}$$

So the ‘Residual’ for the point corresponding to the North East is -0.4.

Similarly the point labelled ‘East’ has coordinates $(493.4, 10.9)$, and its ‘Fit’ value from the scatterplot is 10.6. So for the East,

$$\begin{aligned}\text{Residual} &= \text{Data} - \text{Fit}\\ &= 10.9 - 10.6\\ &= +0.3.\end{aligned}$$

The point representing the North East lies below the line, so its residual is negative. The point representing the East lies above the line, so its residual is positive.

Activity 13 Reading fit values and residuals from a scatterplot

A table of all the data plotted in Figure 30 is given in Table 4. Find the fit values and residuals for the rest of the points on the scatterplot in Figure 30. (The fit values and residuals for the North East and the East have already been entered in the table. These were obtained in Example 9.)

Table 4 Weekly household expenditure and percentage spent on food and non-alcoholic drink

Region	x	y	Fit	Residual
England				
North East	372.7	12.0	12.4	-0.4
North West	430.5	11.4		
Yorkshire and the Humber	405.5	11.4		
East Midlands	449.4	11.6		
West Midlands	430.1	11.7		
East	493.4	10.9	10.6	$+0.3$
London	577.8	9.7		
South East	523.8	10.6		
South West	482.6	11.3		
Wales	394.0	13.1		
Scotland	447.2	11.4		
Northern Ireland	482.8	11.8		

You will have found it quite difficult to measure the fit values accurately using the scatterplot (Figure 30) printed here, as the scale is very small. There is a better method, which involves using the equation of the (fitted) line.

Calculating fit values

The equation of a straight line has the form

$$y = a + bx,$$

where b is the slope or gradient of the line, and a is its intercept (the value of y when $x = 0$). So, for a point where the value of the explanatory variable is x,

$$\text{Fit} = a + bx.$$

You may also see the equation of a straight line written down in other forms, for example as $y = mx + c$ or $y = ax + b$. All these forms are essentially the same, with just changes to the letters representing the slope and the intercept.

Example 10 Calculating a fit value using the equation of the line

It turns out that the line drawn on Figure 30 is given by the equation $y = 18.0 - 0.015x$.

For the point representing the North East, $x = 372.7$. The fit value is therefore

$$\begin{aligned} 18.0 - 0.015x &= 18.0 - 0.015 \times 372.7 \\ &= 18.0 - 5.5905 \\ &= 12.4095, \end{aligned}$$

which is 12.4 when rounded to one decimal place. This is the same as the value that we obtained in Example 9 by reading directly from the graph. In general the two values may not turn out to be exactly the same, because of inaccuracies in reading the graph, and rounding of the calculated value.

Activity 14 Calculating fitted values using the equation of the line

Calculate the fit value for the following situations.

(a) When the equation of the line is $y = 2 + 4x$ and $x = 12$.

(b) When the equation of the line is $y = -4.6 + 0.3x$ and $x = 3$.

(c) When the equation of the line is $y = -0.5x$ and $x = -2.5$.

(d) When the equation of the line is $y = -3.16 - 4.2x$ and $x = -2.7$.

As has already been noted, using the equation of the fitted line allows fit values to be obtained more accurately than by reading them off from a scatterplot. Also, residual values can then be calculated from the fit values and data values using the DFR equation. This determines their values more accurately than by measuring them directly from the scatterplot.

Activity 15 Calculating residuals using the equation of the line

In Example 7 (Subsection 3.1) a straight line was fitted by eye to the data on male unemployment and percentage of households without a car. The equation of this fitted line is $y = 5.8 + 4.2x$, where x is the percentage of males unemployed in a town and y is the percentage of households with no car.

Using the equation of the line, find the residuals for all of the points. For convenience, the data are repeated below.

Table 5 Male unemployment and car ownership for ten towns in England

Town	x	y	Fit	Residual
Alnwick	4.59	21.6		
Vale Royal	3.55	17.2		
Rotherham	5.19	29.7		
Rutland	1.75	13.6		
Dudley	5.27	25.3		
Norwich	5.61	35.5		
Bracknell Forest	2.25	14.5		
Rother	3.00	20.8		
Mole Valley	1.84	13.1		
West Dorset	2.14	16.9		

In this subsection you have learned how to obtain residuals from a line on a scatterplot. In the next subsection you will see how the residuals can be used to help decide whether a line provides a good fit to data points.

You have now covered the material related to Screencast 2 for Unit 5 (see the M140 website).

3.3 Looking for patterns in residuals

We draw a summary line on a scatterplot to try to capture the pattern in the data. If the line is a good fit it should explain all the pattern in the data, and remaining variation around the line should be just random variation. This implies that there should be no pattern in the residuals. If the fit is not a good one, then there may well be some pattern in the residuals.

Example 11 Looking at a set of residuals: 1

In Activities 13 and 15, in the previous subsection, you obtained residuals when the line $y = 5.8 + 4.2x$ was fitted to the data on male unemployment and percentage of households without a car. The residuals, ordered by the male unemployment rate, are given again in Table 6.

Table 6 Residuals from a line fitted to the male unemployment and car ownership data

Town	x	Residual
Rutland	1.75	$+0.4$
Mole Valley	1.84	-0.4
West Dorset	2.14	$+2.1$
Bracknell Forest	2.25	-0.8
Rother	3.00	$+2.4$
Vale Royal	3.55	-3.5
Alnwick	4.59	-3.5
Rotherham	5.19	$+2.1$
Dudley	5.27	-2.6
Norwich	5.61	$+6.1$

Notice that the residuals appear to be centered around zero. This is what we expect for a line that fits the data reasonably well. Since the fit line should represent the overall pattern, we should expect some of the points to be above the fit line and some to be below it. That is, we should expect some of the residuals to be positive and others to be negative.

Example 12 Looking at a set of residuals: 2

Now let's look at the residuals when the line $y = 4.2 + 3.7x$ is fitted to the data on male unemployment and percentage of households without a car. The residuals from fitting this line are given in Table 7.

Table 7 Residuals from a different line fitted to the male unemployment and car ownership data

Town	x	y	Fit	Residual
Rutland	1.75	13.6	10.7	$+\ 2.9$
Mole Valley	1.84	13.1	11.0	$+\ 2.1$
West Dorset	2.14	16.9	12.1	$+\ 4.8$
Bracknell Forest	2.25	14.5	12.5	$+\ 2.0$
Rother	3.00	20.8	15.3	$+\ 5.5$
Vale Royal	3.55	17.2	17.3	$-\ 0.1$
Alnwick	4.59	21.6	21.2	$+\ 0.4$
Rotherham	5.19	29.7	23.4	$+\ 6.3$
Dudley	5.27	25.3	23.7	$+\ 1.6$
Norwich	5.61	35.5	25.0	$+10.5$

Notice that all but one of the residuals are positive. This pattern suggests that the fit values are generally too low. Hence, the line should be higher if it is going to fit the data reasonably well.

Example 13 Looking at a set of residuals: 3

Some data on oxygen uptake were introduced in Activity 6 (Subsection 2.1). A straight line was fitted to these data and the resulting residuals are given in Table 8. (In the table, only every fifth residual is given to make the table simpler.)

Table 8 Residuals from the oxygen-uptake experiment

Oxygen uptake	Residual
667	$+\ 4.7$
1020	$-\ 3.4$
1599	-10.9
1874	-11.4
2312	-12.3
2766	$-\ 6.8$
3151	$-\ 2.3$
3521	$+\ 6.0$
3878	$+14.8$
4290	$+58.7$

You can see that the residuals are positive for both small and large values of oxygen uptake (the explanatory variable) and they are negative for intermediate values.

This has a pattern to it. It is a pattern that is related to the values of the explanatory variable. This suggests that we could look for a relationship between the residuals and the values of the explanatory variable – and hence look for a relationship between the response variable and the explanatory variable which is over and above that explained by the fit line. So the fit line does not capture all of the relationship between the response variable and the explanatory variable.

The conclusion at the end of Example 13 holds in general.

Residual patterns

If the residuals show a pattern that relates to the explanatory variable, then the fit line does not provide an adequate explanation of all the patterns in the data, and we should look for a better relationship.

You have now covered the material related to Screencast 3 for Unit 5 (see the M140 website).

Some patterns in residuals are easy to spot by looking at a table. However, it is more usual to investigate a possible relationship between the residuals and the values of the explanatory variable using a scatterplot in which the horizontal coordinate is the explanatory variable, exactly as in the original scatterplot, and the vertical coordinate is the residual. Such a scatterplot is called a **residual plot** or sometimes a residual scatterplot.

Example 14 A residual plot

Figure 31(a) shows a scatterplot of the oxygen uptake and expired ventilation data from Figure 6, together with a fit line. (To make the plot simpler, only every fifth data point is shown – similarly to Table 8.) Figure 31(b) shows the corresponding residual plot. On the residual plot, notice a line corresponding to a residual value of zero – shown for reference.

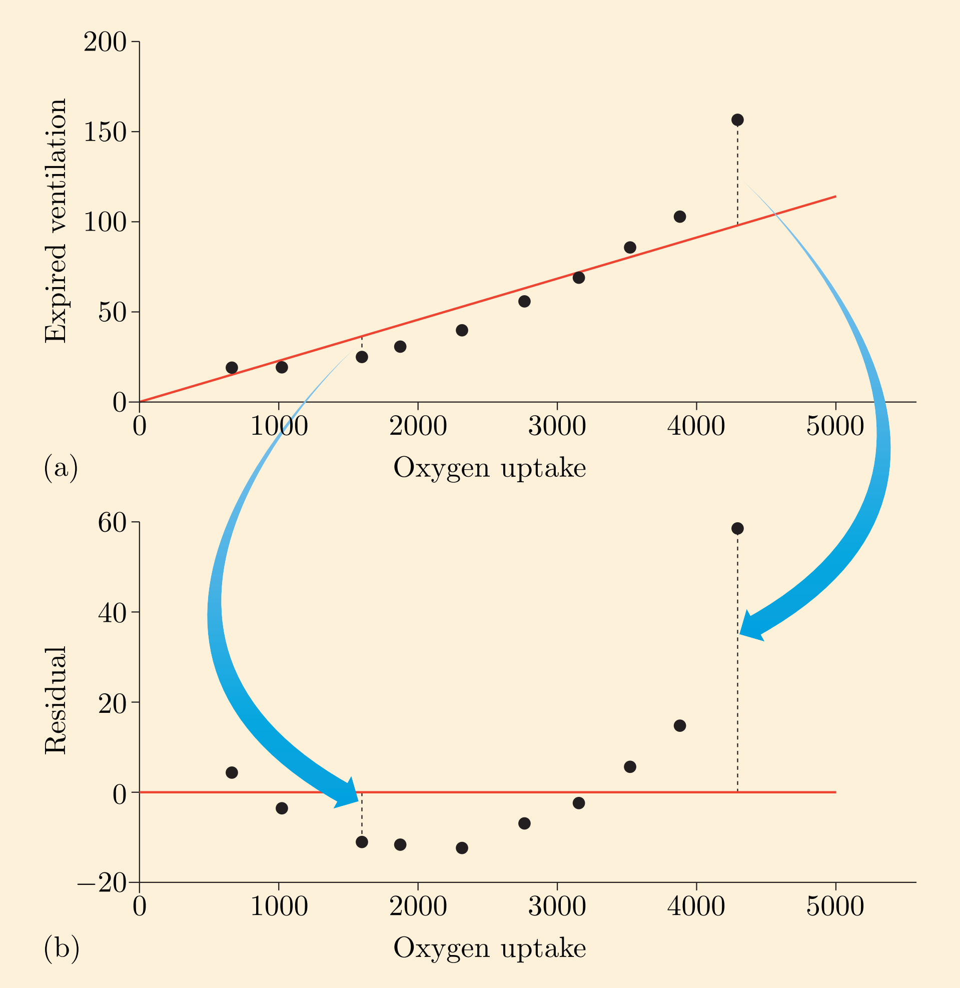


Figure 31 Data from an experiment in kinesiology, with (a) a straight fit line and (b) showing residuals

The residual plot shows a definite curved pattern. A straight line does not provide a satisfactory fit to the data. The residuals actually look larger in the residual plot than they do in the scatterplot, because the vertical scale has been increased to show the residuals more clearly. In this example, the curve could be seen in the original scatterplot, but often patterns are easier to spot in a residual plot.

Example 14 is the subject of Screencast 4 for Unit 5 (see the M140 website).

Activity 16 Working with a residual plot

An experiment was carried out at Charing Cross Hospital on the effect of the drug captopril (Figure 32) on the blood pressure of patients with moderate essential hypertension. The diastolic blood pressure of 15 patients was measured immediately before, and two hours after, receiving an injection of the drug. (Source: MacGregor, Markandu, Roulston and Jones (1979), *British Medical Journal*, vol. 2, pp. 1106–1109)

(Note that 'moderate essential hypertension' is a disorder involving blood pressure and 'diastolic blood pressure' is the lowest pressure between heartbeats.)

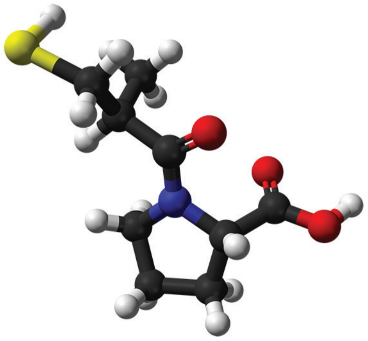


Figure 32 A model of the drug captopril

The results of the experiment are shown in the scatterplot in Figure 33 and a straight line has been fitted by eye to this data. (Obviously, blood pressure *before* treatment must be the explanatory variable, as this can influence blood pressure after treatment, but this cannot be true the other way round.)

Figure 34 shows four possible residual plots for these data and the line.

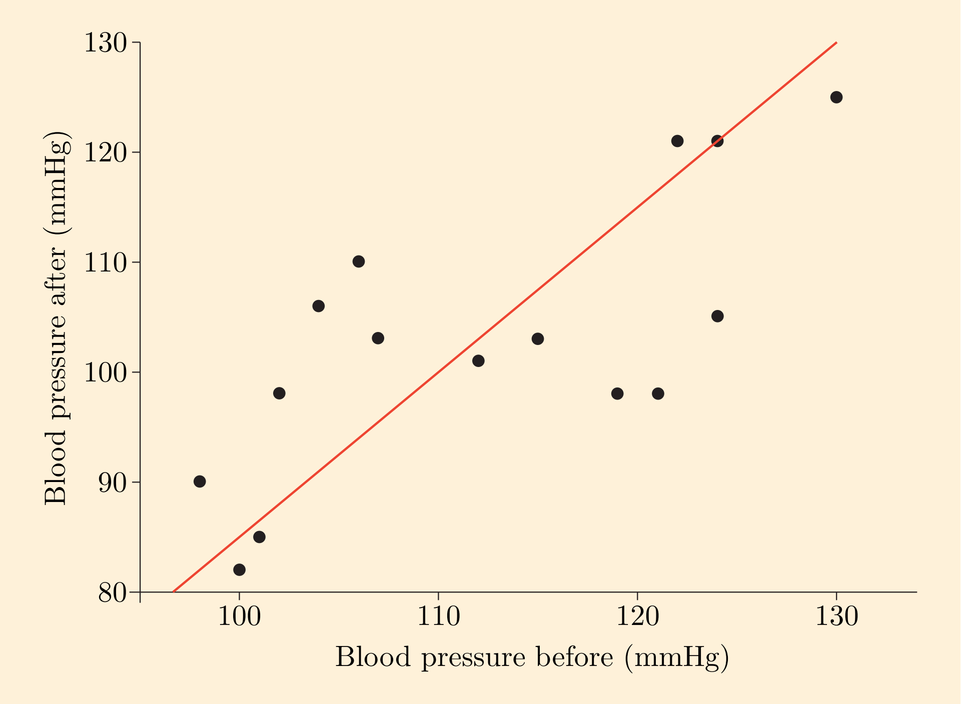


Figure 33 Blood pressure data from captopril study

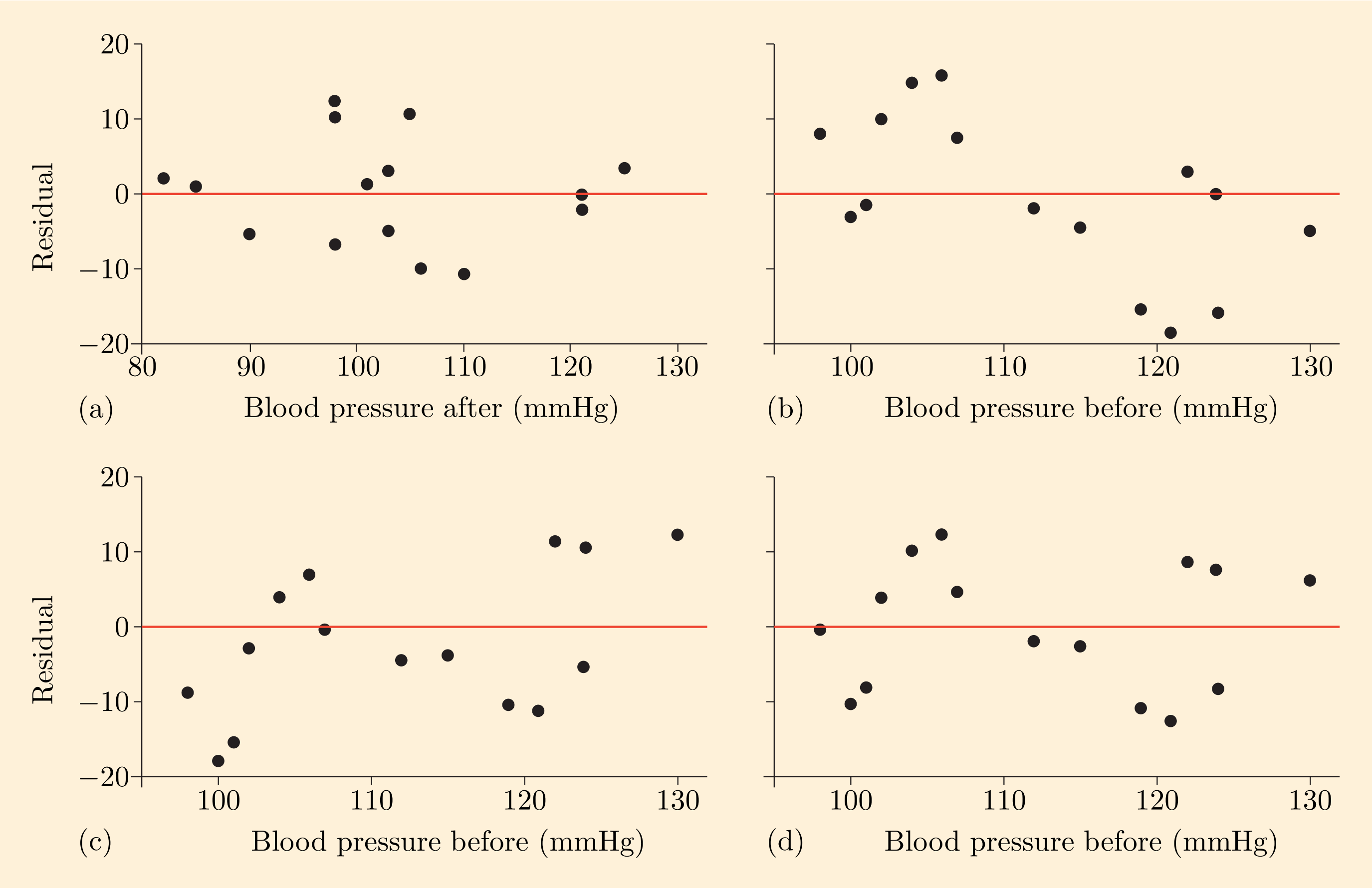


Figure 34 Four possible residual plots for the blood pressure data

(a) Which one of the four residual plots in Figure 34 corresponds to the correct residual plot for the line shown in Figure 33?

(b) Can you spot a pattern in the correct residual plot from Figure 34? If so, how should the fit line in Figure 33 be moved?

In Activity 16, the residuals suggest that the line fitted by eye is not the best that could be drawn and that a line which is a little less steep would be better. We could proceed by drawing another line, finding the residual values and drawing another residual plot. But this is rather a hit-and-miss procedure. We might find that it is now not steep enough, or perhaps that it was a little too high so that too many residuals were negative. Also, no two people would end up drawing exactly the same straight line. Moreover, the procedure is very tedious, particularly if there are a lot of points. In Section 4 you will learn a method of calculating the equation of a straight line that provides a good fit to a set of data.

Exercises on Section 3

Exercise 5 Summarising change in house prices over time

Exercise 4 featured a scatterplot of average prices in the UK between 1991 and 2008. This scatterplot is repeated below for convenience.

Use this scatterplot to add a line that you feel provides a good summary of the data.

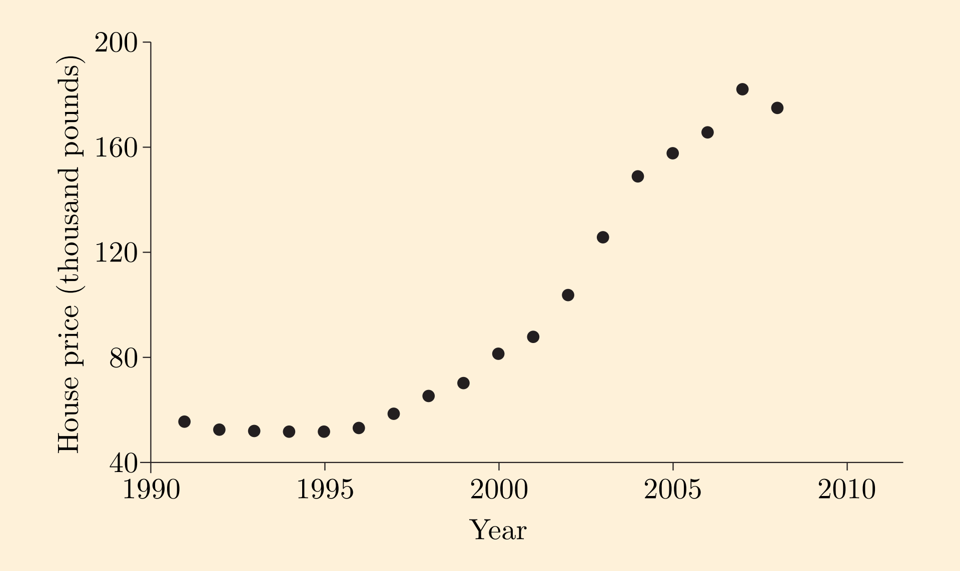


Figure 35 Average house price in the UK between 1991 and 2008

Exercise 6 Calculating fitted values and residuals

For a data point where $x = 20$ and $y = 4$, calculate the following.

(a) The fit value when the equation of the line is $y = 125 - 6x$.

(b) The fit value when the equation of the line is $y = -3 + 0.25x$.

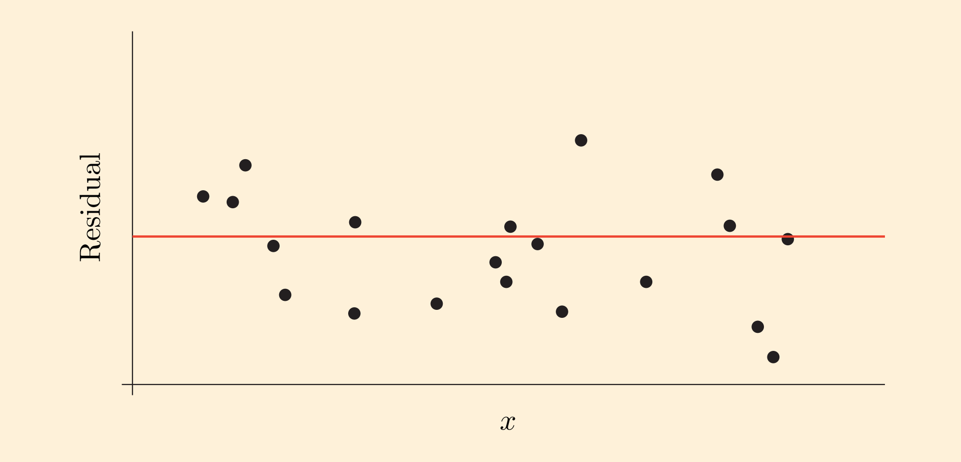
(c) The residual when the equation of the line is $y = 0.15x$.

(d) The residual when the equation of the line is $y = 8 + x$.

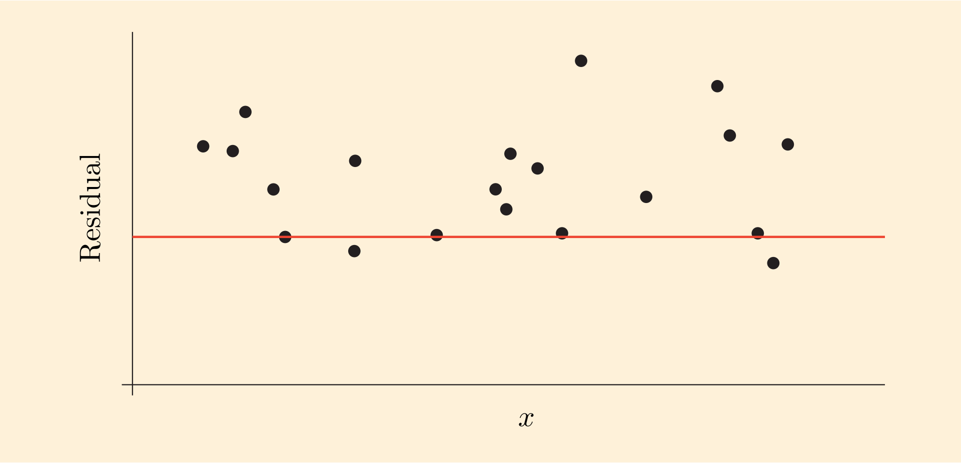
Exercise 7 Assessing the fit of lines

Four different lines have been fitted to a set of 20 data points. The corresponding residual plots are shown below. Also shown on each plot is the line corresponding to a residual value of zero. Using each plot, comment on the fit of the line. If you think that the line does not fit very well, suggest how the line should be moved so that it fits the data better.

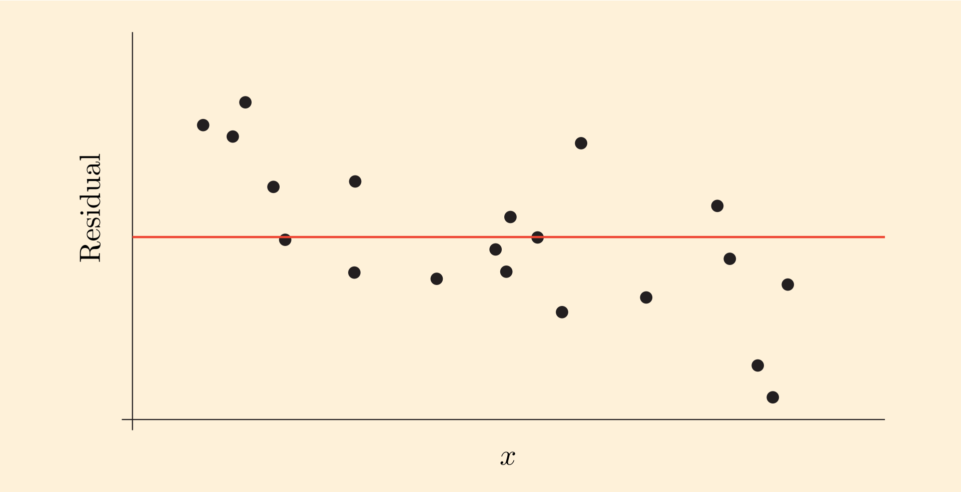
(a)



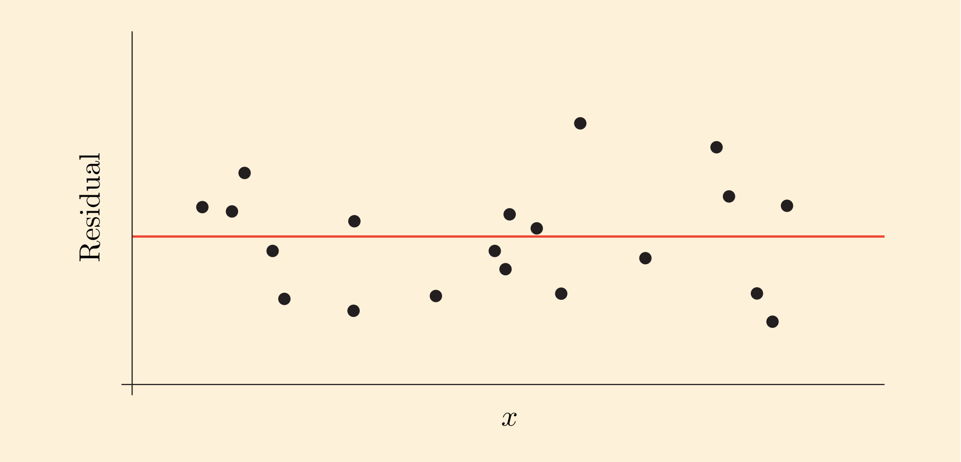
(b)



(c)



(d)



4 The least squares regression line

In the last section, we concentrated on fitting lines by eye. This is a simple, straightforward method and is often an adequate approach when exploring data to get some idea of what relationship is present. However, fitting lines by eye is subjective – different people draw different lines. If you are going to report your investigation to someone else, there is a need for an objective, well-defined procedure for drawing a straight fit line on a scatterplot. If you have used a formal method, you can describe exactly what you did, and another person presented with the same dataset would get exactly the same line by using your method.

Another reason for requiring a formal method is that when computers are to be used to fit lines, they have to be instructed exactly how to carry this out; a computer cannot just draw a line that 'appears' to be a good fit! Also, it is good to choose a straight line that is optimal in some way.

Several formal methods of fitting lines exist. In this section, we introduce what is by far the commonest of these, which is known as fitting by the method of **least squares**. It has many useful properties, some of which will be discussed here. The resulting line is called the **least squares fit line** or the **least squares regression line**.

4.1 What is least squares?

The method of least squares is based on a study of residuals obtained when different lines are fitted to a set of data.

A line is a good fit to a batch of data if the residuals are small. When all of the points lie exactly on a straight line, as in Example 4 (Subsection 2.2), all the residuals are zero and the fit is perfect. However, this is very rarely the case in practice, and so we need a method that chooses a line for which the residuals are as small as possible.

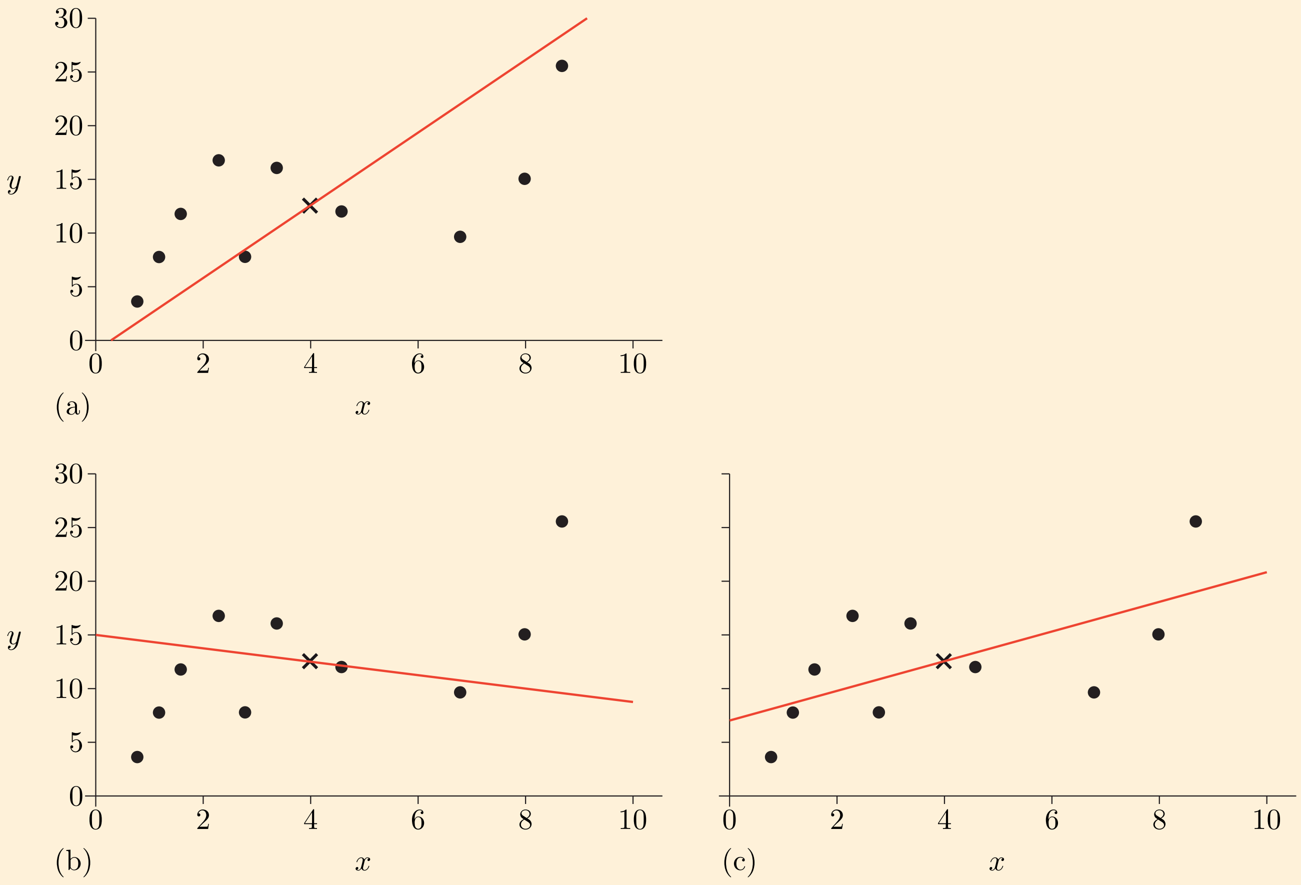
In Example 12 (Subsection 3.3) all but one of the residuals were positive and we suggested that the line was too low. Instinctively, a good line should be somewhere in the middle of the data. The method of least squares takes care of this.

The method of least squares

This method is used to find a good fit line, by choosing a line that passes through the overall mean of the data: the line goes through the point whose x-coordinate is the mean of the x-values in the data and whose y-coordinate is the mean of the y-values in the data. This point can be denoted $(\overline{x}, \overline{y})$, where $\overline{x}$ and $\overline{y}$ are the two means.

It can be shown that if a line passes through the point $(\overline{x}, \overline{y})$, then the sum of all the residual values, taking their signs into account, is always zero. In other words, the total of all the positive residuals is equal to the total of all the negative residuals. You do not have to worry about why this is so, but you can see it is a useful property of the fit line; it ensures that there are not too many positive or too many negative residuals. It also takes account of the situation where we might have one or two large positive residuals and all the other residuals are small and negative. It is unnecessary for there to be equal numbers of positive and negative residuals.

Requiring the fit line to go through $(\overline{x}, \overline{y})$ only gives one point that the line must pass through; something else is needed to choose the best slope. Look at Figure 4.1.



This figure shows three possible fit lines for the same set of data points. All three lines pass through the point $(\overline{x}, \overline{y})$, which is marked with a cross. The line in Figure 4.1(a) is too steep, and the line in Figure 4.1(b) is not steep enough. In both of these, the lengths of some of the residuals, ignoring their signs, are quite large. By contrast, the lengths of all the residuals in Figure 4.1(c) are small, and this line is the best fit of the three lines illustrated.

Residual plots for the lines shown in Figure 4.1(b) and (c) have been drawn in Figure 36(a) and (b). Look just at the *lengths* of the residuals. You can see that for almost every point the residual is shorter in Figure 36(b) than in (a). This is what you expect: the line in Figure 4.1(c) looks to be a better fit than the line in Figure 4.1(b) *because* it gives shorter residuals.

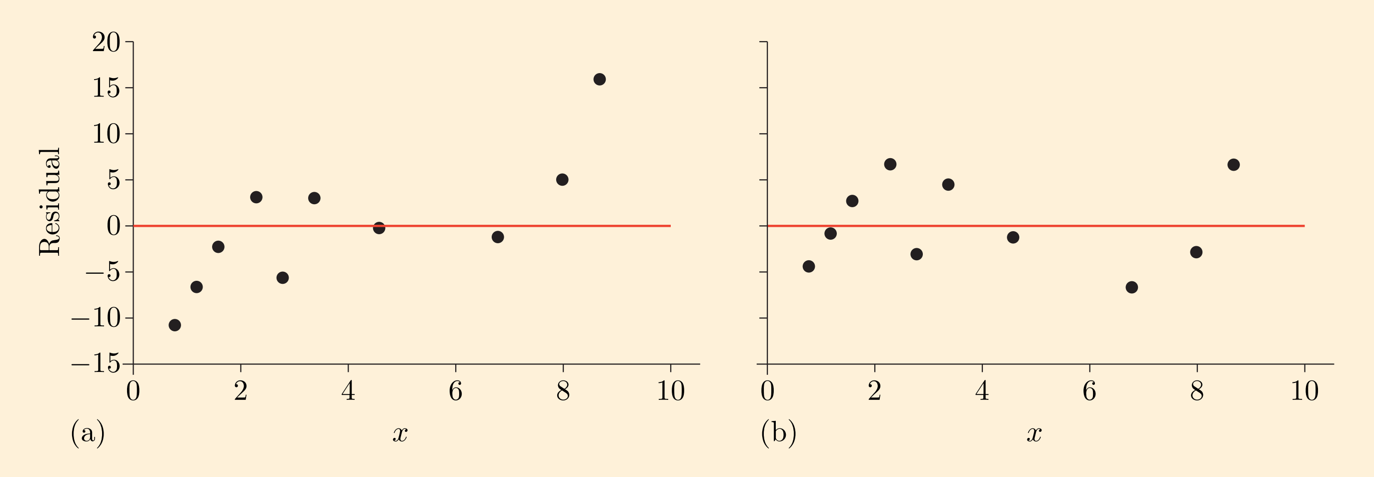


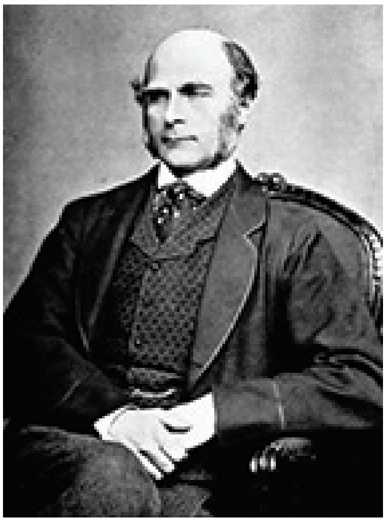
Figure 36 Residual plots corresponding to Figure 4.1(b) and (c)

The method of least squares uses the fact that lines that give a good fit have short residuals. To get rid of the negative signs, the lengths of all residuals are squared. These squares are all added together, and the slope is chosen so as to make this sum of squared residuals as small as possible. It is this property that gives the method its name – *least squares*.

The least squares regression line

The least squares regression line is the line for which the sum of the squares of the residuals is minimised.

The least squares regression line always goes through the point $(\overline{x}, \overline{y})$ and hence the sum of the residuals is always zero.



Sir Francis Galton

Sir Francis Galton and regression

Sir Francis Galton (1822–1911) made important contributions in many fields. He was obsessed with data and measured everything that he could, from wind direction (he was an initiator of scientific meteorology) to fingerprints. The latter led to him devising a method of classifying fingerprints that proved useful in forensic science. He was also a geographer and explorer in his early years (notably in south-west Africa) and founded the science of measuring mental faculties (psychometrics). However, he is best known for his work in anthropology, heredity and eugenics, and statistics.

He was a half-cousin of Charles Darwin and was greatly influenced by Darwin's book *Origin of Species*, published in 1859. Following its publication, much of Galton's work focused on exploring variation in human populations and on whether human ability was inherited or learned – he coined the phrase *nature or nurture*. To better understand the large quantities of data that he collected, he devised and extended a number of statistical techniques, including regression.

One dataset he collected were the heights of 205 parents and their 930 adult children. The heights of women were multiplied by 1.08, so as to adjust for gender. Galton then plotted the height of a child (the response) against the average height of its parents (the explanatory variable) and represented their relationship by a straight line. Galton noted that the children of the shorter parents tended to be taller than their parents, and the children of the taller parents tended to be shorter than their parents. (Source: Galton, F. (1886) 'Regression towards mediocrity in hereditary stature', *The Journal of Anthropological Institute of Great Britain and Northern Ireland*, vol. 15, pp. 246–263.)

This regression of a child's measurement towards the mean value in the population is a characteristic of inherited attributes. It is called *regression to the mean*. This example underlies the use of *regression* in the phrase *least squares regression*.

You have now covered the material needed for Subsection 5.1 of the Computer Book.

4.2 Calculating the least squares regression line by hand

Calculating the least squares regression line bears many similarities to calculating the standard deviation for a single variable, in particular the calculation using Method 2. So you may find it helpful to revise Subsection 3.1 of Unit 3 before studying this subsection.

The least squares regression line has the form

$$y = a + bx.$$

We must calculate its slope, b, and its intercept, a. It turns out that the formula for the slope of the least squares regression line is as follows:

$$b = \frac{\sum(y-\overline{y})(x-\overline{x})}{\sum(x-\overline{x})^2}$$

The summation notation was first introduced in Subsection 1.3 of Unit 2.

and that the formula for the intercept of the least squares regression line is

$$a = \overline{y} - b \times \overline{x}.$$

These formulas look daunting at first, and how they are derived from the definition of the least squares regression line is beyond the scope of M140. So the method for calculating the regression line using least squares will be demonstrated in this subsection by means of an example. We will calculate the regression line for the data on percentages of males unemployed and households with no car.

The calculation will be broken down into five steps. The first step is to calculate the sum of all the x-values, the sum of all the y-values, the sum of the squares of all the x-values and the sum of the products of the x- and y-values. That is, we will calculate

$$\sum x,\ \sum y,\ \sum x^2 \text{ and } \sum xy.$$

Example 15 Calculating a least squares regression line – step 1

In the scatterplots of these data in this unit, the percentage of men unemployed plotted along the x-axis, and the percentage of households with no car along the y-axis. So in terms of x and y, the data are as follows.

Table 9 Male unemployment and car ownership for ten towns in England

Town	x	y
Alnwick	4.59	21.6
Vale Royal	3.55	17.2
Rotherham	5.19	29.7
Rutland	1.75	13.6
Dudley	5.27	25.3
Norwich	5.61	35.5
Bracknell Forest	2.25	14.5
Rother	3.00	20.8
Mole Valley	1.84	13.1
West Dorset	2.14	16.9

The sums of all the x-values and of all the y-values are as follows:

$$\sum x = 4.59 + \cdots + 2.14 = 35.19, \qquad \sum y = 21.6 + \cdots + 16.9 = 208.2.$$

We also require the sum of the squares of the x-values and the sum of the products of the x- and y-values. You should be able to find these two sums on your calculator without writing down each square (or product) separately.

$$\begin{aligned}\sum x^2 &= 4.59^2 + 3.55^2 + \cdots + 2.14^2\\ &= 144.9419,\\ \sum xy &= 4.59 \times 21.6 + 3.55 \times 17.2 + \cdots + 2.14 \times 16.9\\ &= 825.928.\end{aligned}$$

These four sums are the basic quantities you need, and this completes the first step of the calculations.

The second step is to calculate the means of the x- and y-values. The third step is to calculate the sum of the squared deviations of the x-values and the sum of the products of the deviations of the x- and y-values. That is, in steps 2 and 3, we will calculate

$$\overline{x},\ \overline{y},\ \sum(x-\overline{x})^2 \text{ and } \sum(x-\overline{x})(y-\overline{y}).$$

Example 16 Calculating a least squares regression line – steps 2 and 3

Step 2 is the calculation of means of the x and y values.

In this dataset there are ten observations, so $n = 10$.

The mean of x is therefore $\overline{x} = 35.19/10 = 3.519$, and the mean of y is therefore $\overline{y} = 208.2/10 = 20.82$.

In step 3, the sum of the squared deviations of the x-values is calculated. This sum is one that you have encountered before, as part of the calculation of a standard deviation. Here part of 'Method 2' for the standard deviation is used to

The calculation of the standard deviation was introduced in Subsection 3.1 of Unit 3.

calculate $\sum(x-\overline{x})^2$.

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n}\\ &= 144.9419 - \frac{(35.19)^2}{10}\\ &= 144.9419 - 123.833\,61 = 21.108\,29.\end{aligned}$$

The sum of the products of the deviations of the x- and y-values can be calculated in a similar way using the sum of the products of the x- and y-values along with the mean of the x-values and the mean of the y-values.

$$\begin{aligned}\sum(x-\overline{x})(y-\overline{y}) &= \sum xy - \frac{(\sum x)(\sum y)}{n}\\ &= 825.928 - \frac{35.19\times 208.2}{10}\\ &= 825.928 - 732.6558 = 93.2722.\end{aligned}$$

The final two steps involve calculating the slope and the intercept of the regression line. The only terms required are those calculated in steps 2 and 3.

Example 17 Calculating a least squares regression line – steps 4 and 5

The slope, b, of the regression line is given by the following formula.

$$b = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2}.$$

So in this example

$$b = \frac{93.2722}{21.108\,29} \simeq 4.418\,747,$$

which we round to three significant figures as 4.42 (the same number of significant figures as the percentage of men unemployed is given to).

The intercept, a, of the regression line is given by:

$$a = \overline{y} - b\times\overline{x}.$$

So in this example

$$a \simeq 20.82 - 4.418\,747\times 3.519 = 5.270\,429\,307,$$

which we round to two decimal places as 5.27, one more decimal place than that used for the percentage of households without a car.

The equation of the least squares regression line is therefore

$$y = 5.27 + 4.42x.$$

Often the least squares regression line is referred to as simply ‘the regression line’, and calculating and using the line is referred to as **linear regression**. The procedure for its calculation can be summarised as follows.

Calculating the least squares regression line $y = a + bx$ for a set of n data points (x, y)

1. Calculate $\sum x$, $\sum y$, $\sum x^2$ and $\sum xy$.

2. Calculate the means of x and y:

$$\overline{x} = \frac{\sum x}{n} \quad \text{and} \quad \overline{y} = \frac{\sum y}{n}.$$

3. Calculate the sum of the squared deviations of the x-values

$$\sum(x-\overline{x})^2 = \sum x^2 - \frac{\sum x^2}{n},$$

and the sum of the products of the deviations

$$\sum(x-\overline{x})(y-\overline{y}) = \sum xy - \frac{\sum x \sum y}{n}.$$

4. The slope b is given by

$$b = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2}.$$

5. The intercept a is given by

$$a = \overline{y} - b\overline{x}.$$

You have now covered the material related to Screencast 5 for Unit 5 (see the M140 website).

To draw the fitted line on a scatterplot, we just calculate the coordinates of two well-separated points on the line and then draw a straight line through them. After drawing the line you should look at it to check that it seems right – it should appear to pass through the middle of the data. If it clearly does not, then there is a calculation error and you should check your working.

Example 18 Drawing the regression line on a scatterplot

In Example 17, the regression line was calculated as

$$y = 5.27 + 4.42x.$$

From Table 9 in Example 15, the scatterplot, the x-values of the data range from about 2 to 6. We substitute these values into the equation of the regression line to obtain the coordinates of two well-separated points.

When $x = 2$,

$$y = 5.27 + 4.42 \times 2 = 14.11,$$

so one point on the line is (2, 14.11).

When $x = 6$,

$$y = 5.27 + 4.42 \times 6 = 31.79,$$

so a second point on the line is (6, 31.79).

Figure 37 shows the line drawn on the scatterplot. You can see that it appears to provide a reasonably good fit to the points. If you compare it to Figure 23 (Subsection 3.1), you can see that the line fitted by least squares is slightly steeper than the one drawn by eye.

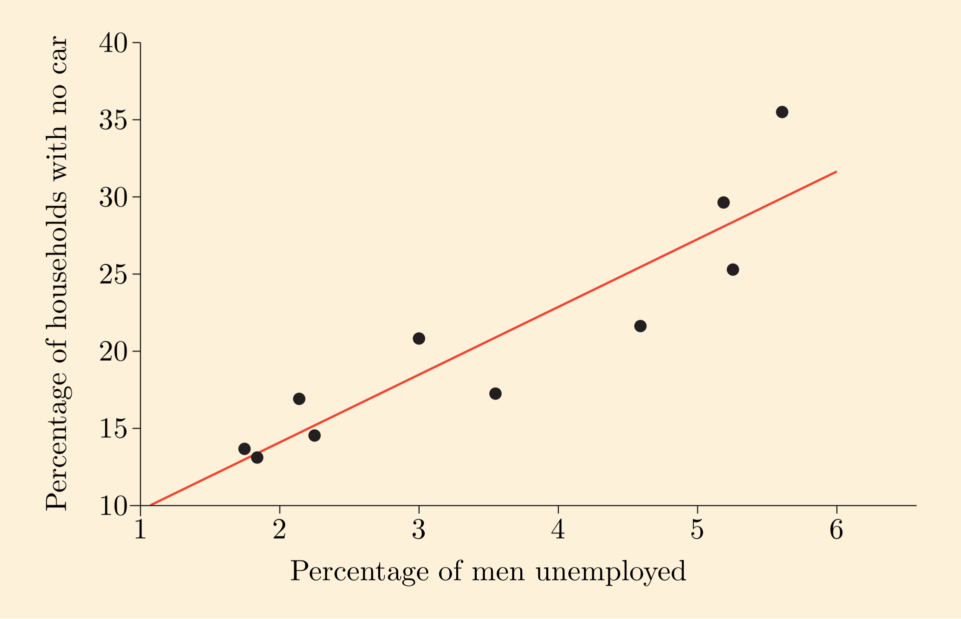


Figure 37 Percentage of males unemployed and percentage of households with no car, with least squares line

The next activity provides practice in calculating a regression line.

Activity 17 Calculating a least squares regression line

Activity 16 (Subsection 3.3) featured a line fitted by eye to diastolic blood pressure data from a study of the drug captopril. The data shown in Figure 33 of Activity 16 are given in Table 10. Using these data, calculate the least squares regression line. Also, add the regression line to Figure 38.

Table 10 Diastolic blood pressure before and after injection

x, blood pressure before injection (mmHg)	y, blood pressure after injection (mmHg)
130	125
122	121
124	121
104	106
112	101
101	85
121	98
124	105
115	103
102	98
98	90
119	98
106	110
107	103
100	82

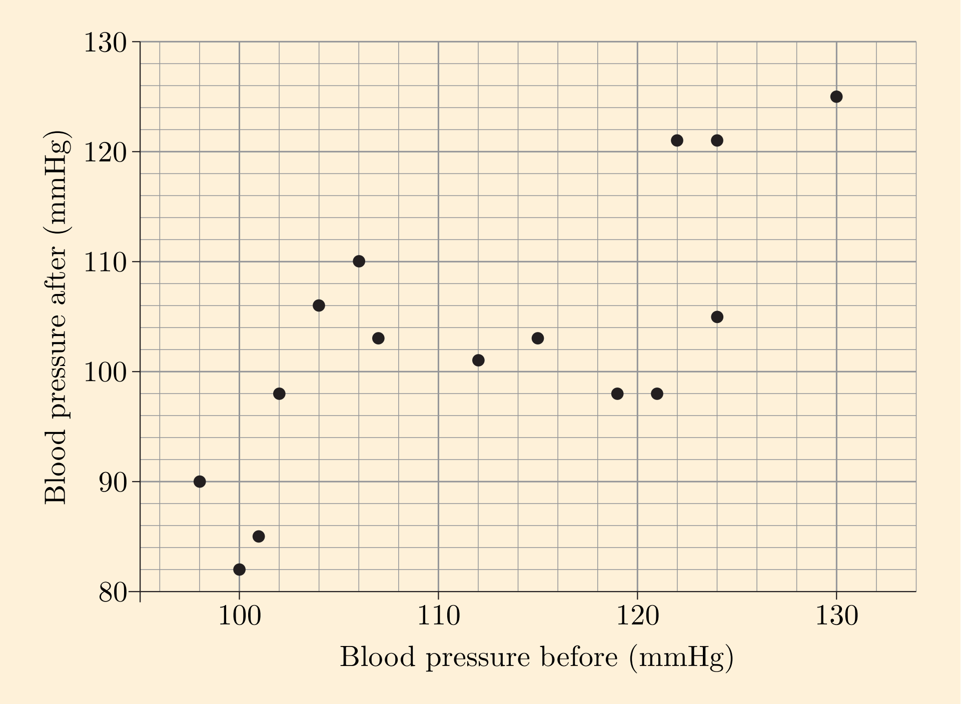


Figure 38 Scatterplot of blood pressure data from captopril study

Exercises on Section 4

Exercise 8 Spotting the least squares regression line

In Figure 39, three lines fitted to a set of 12 data points are shown. One of these lines is the least squares regression line. Identify which one it is. For each of the two lines that are not the least squares regression line, give a reason why it is not.

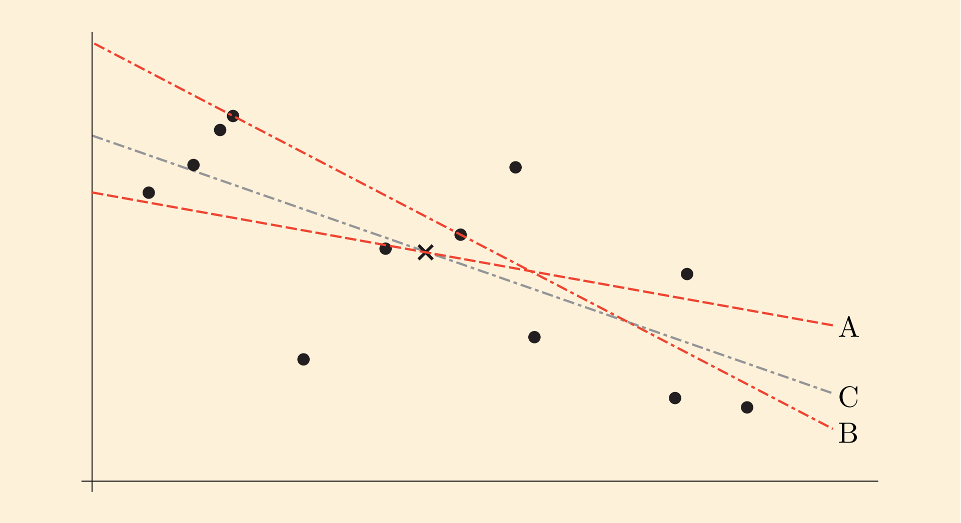


Figure 39 Scatterplot of some data along with three fitted lines. (The point $(\overline{x}, \overline{y})$ is also marked on the scatterplot.)

Exercise 9 Fitting a line to house prices

In the 1980s, the average UK house prices were as follows. Using these data, calculate the least squares regression line.

Table 11 Average house prices in the UK

x, Year	y, House price (\$ thousands)
1980	23.3
1981	24.1
1982	24.7
1983	27.4
1984	30.8
1985	34.2
1986	37.0
1987	43.0
1988	48.9
1989	62.2

(Data source: Nationwide building society (2013) ‘UK house prices since 1952’)

5 Using the least squares regression line

The previous section introduced the least squares regression line and showed how it can be calculated. Once we have calculated a regression line for a sample of data points, what can we do with it? In this section we will explore two things: checking that the regression line fits the data well, and using the line for prediction.

‘Data don’t make any sense, we will have to resort to statistics.’

5.1 Checking the least squares regression line

In Section 6 you will see that least squares is a method by which a straight line can be fitted to data automatically. Although the least squares regression line should be as good as any straight line fitted by eye, it is possible that no straight line fits the data well. So, it is important to check that the least squares line provides an adequate fit. This can be done by looking at the residuals.

Since we have used the least squares method to fit the line, there is no need to look for two of the patterns that we sometimes found in Subsection 3.3. The average of the residuals is zero, so there cannot be a pattern of too many positive (or negative) residuals. It is possible that there will be more, say, small positive residuals and fewer large negative residuals, but the least squares method chooses a line where the sum of the residuals is zero. It also chooses the slope so that the sum of the squared residuals is as small as possible. This ensures that there will be no tendency for positive residuals to be associated with, say, small values of x, and negative residuals with large values of x. Otherwise, the line would be rotated about the point $(\overline{x}, \overline{y})$, as we saw in Figure 4.1.

The patterns that we *might* see include, for example, positive residuals for both small and large values of x, and negative residuals for intermediate x, which indicates that a curved line would fit the data better.

Activity 18 Interpreting residual plots

The following three residual plots are the result of fitting least squares lines to three different sets of data. Use each residual plot to state how reasonable a straight-line model is for the dataset. Justify your opinion.

(a)

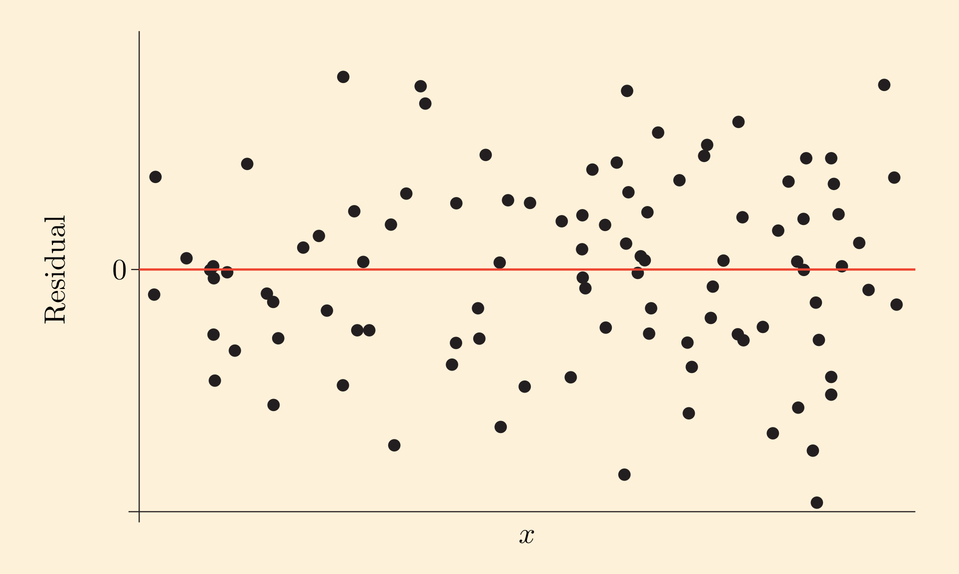


Figure 40 Residual plot for dataset 1

(b)

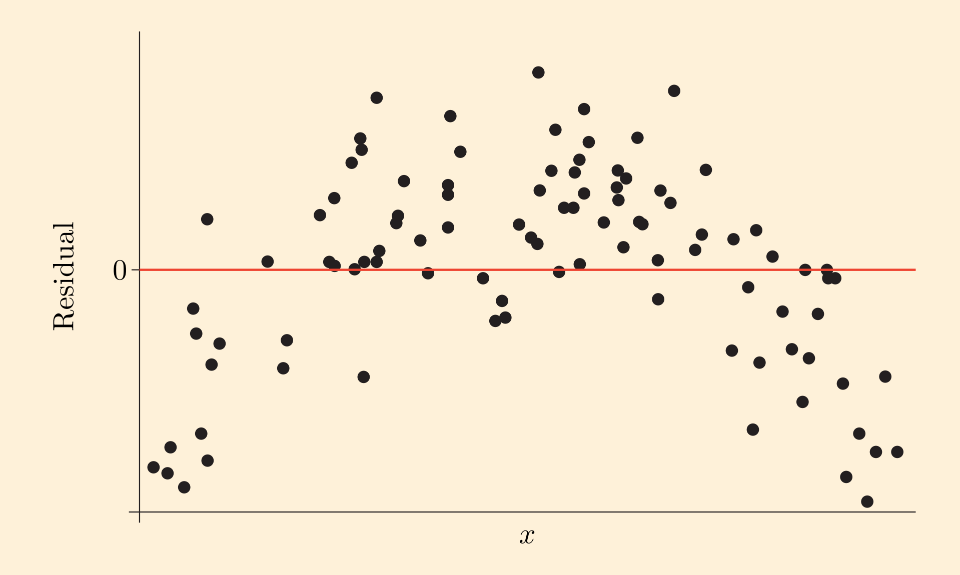


Figure 41 Residual plot for dataset 2

(c)

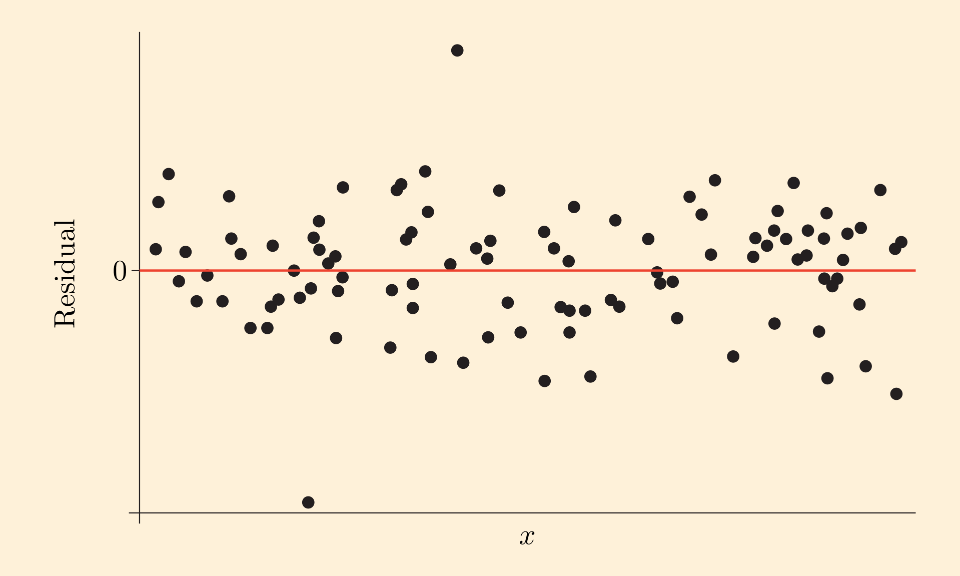


Figure 42 Residual plot for dataset 3

So, when the line fits well there is no pattern in the residual plot. That is the residuals are evenly scattered above and below the line $y = 0$. Possible patterns that indicate problems with the fit include the following:

- A non-linear relationship between residuals and the explanatory variable (which indicates that a curved line will probably be more appropriate).
- A particularly large or small residual. That is, one that does not follow the pattern of the other residuals. This indicates that there is probably an outlier in the data.

Residuals can be found most easily using the equation of the least squares line, following the same process as in Subsection 3.2. If the line has the equation $y = a + bx$, then for each value of x in the sample, we calculate the fit value and then the residual.

$$\text{Fit} = a + bx.$$

$$\begin{aligned}\text{Residual} &= \text{Data} - \text{Fit}\\ &= y - (a + bx).\end{aligned}$$

Let us return to the example on male unemployment and households without cars.

Example 19 Examining residuals for a least squares regression line

The least squares regression line calculated in Subsection 4.2 was $y = 5.27 + 4.42x$. We shall use this line to calculate the residual for the first data point, Alnwick, (4.59, 21.6):

$$\begin{aligned}\text{Fit} &= 5.27 + 4.42 \times 4.59\\ &= 5.27 + 20.2878\\ &= 25.6 \text{ (rounded to one decimal place).}\end{aligned}$$

So,

$$\begin{aligned}\text{Residual } &\simeq 21.6 - 25.6\\ &= -4.0.\end{aligned}$$

Table 12 shows all the fit and residual values (rounded to one decimal place).

Table 12 Residuals and fitted values for ten towns

Town	x	y	Fit	Residual
Alnwick	4.59	21.6	25.6	-4.0
Vale Royal	3.55	17.2	21.0	-3.8
Rotherham	5.19	29.7	28.2	$+1.5$
Rutland	1.75	13.6	13.0	$+0.6$
Dudley	5.27	25.3	28.6	-3.3
Norwich	5.61	35.5	30.1	$+5.4$
Bracknell Forest	2.25	14.5	15.2	-0.7
Rother	3.00	20.8	18.5	$+2.3$
Mole Valley	1.84	13.1	13.4	-0.3
West Dorset	2.14	16.9	14.7	$+2.2$

Figure 43 shows the residual plot for this example. There is no obvious pattern to be seen here, so this suggests that a straight line is a reasonable model for the data points.

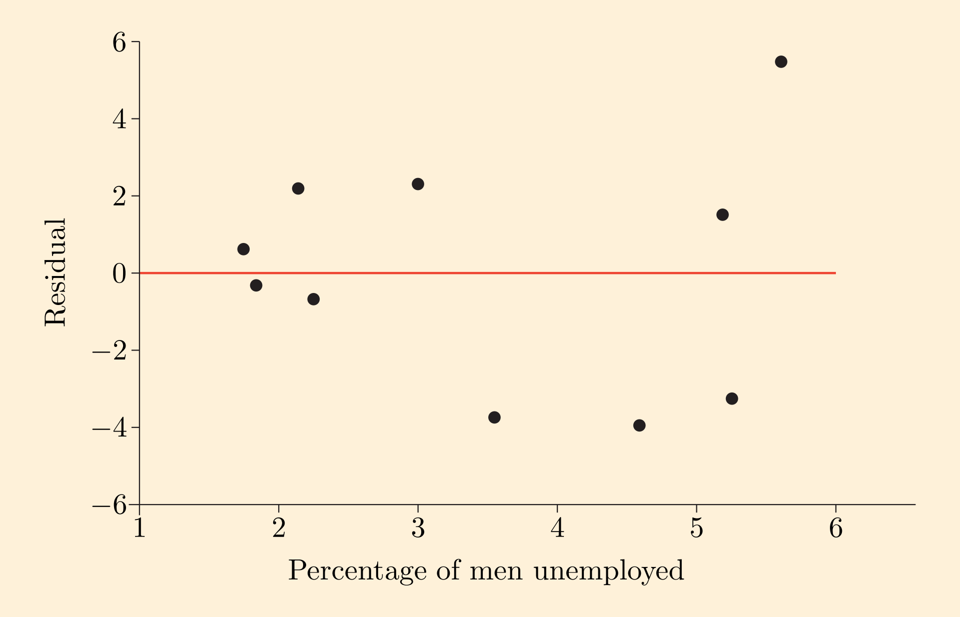


Figure 43 Residual plot for the male unemployment and car ownership data, using the least squares line

Example 19 is the subject of Screencast 6 for Unit 5 (see the M140 website).

Activity 19 Examining the fit of a least squares regression line

In Activity 17 (Subsection 4.2) you calculated the least squares regression line for the blood pressure data from the study involving the drug captopril.

(a) Using the equation of the regression line, $y = 4.2 + 0.880x$, calculate the residuals for the first five observations. For convenience, the data for these observations are given again below.

x	y
130	125
122	121
124	121
104	106
112	101

(b) Figure 44 shows the residual plot for all the data. Comment on the fit of the regression line to these data.

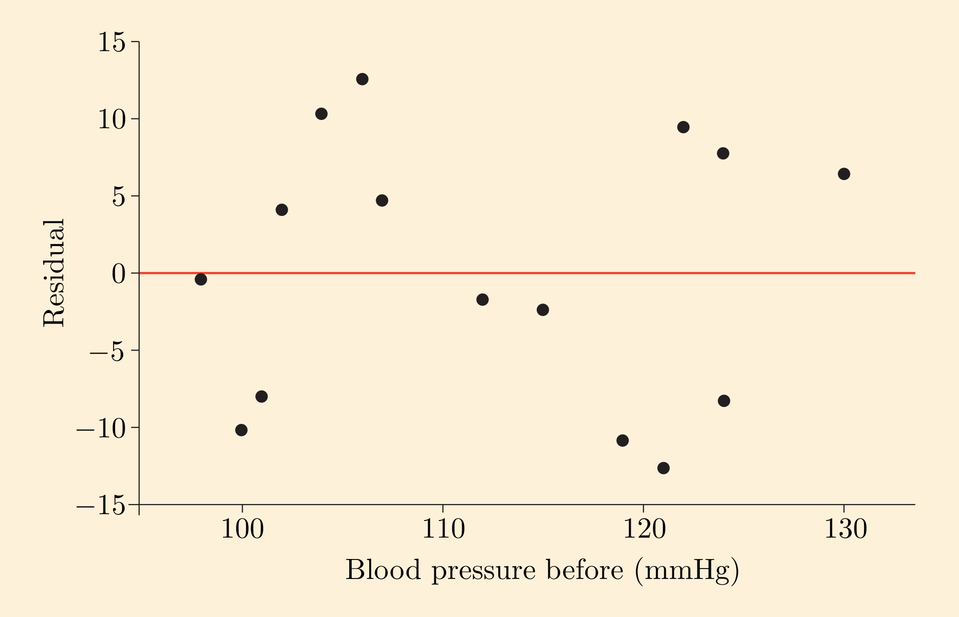


Figure 44 Residual plot for the blood pressure data, using least squares regression

5.2 Using the least squares regression line for prediction

Once we have established that the regression line does provide a reasonable model for a set of data, what can we do with it? One of its uses is for **prediction**. If we know the value of the explanatory variable for some individual, then we can forecast the value of the response variable for that individual.

'Mr Palmer, using statistics, I can predict which numbers will be chosen in the lottery .... I just don't know when.'

In Activity 17 (Subsection 4.2), you calculated the following regression line for diastolic blood pressure before (x) and after (y) injection with captopril for patients with moderate essential hypertension: $y = 4.2 + 0.880x$. Suppose another patient with moderate essential hypertension arrives at the hospital. If a doctor measures the patient's diastolic blood pressure, then by using the equation of the regression, the doctor can calculate the fit value. This value is a prediction of the patient's diastolic blood pressure two hours after injection with captopril. For example, if the patient's blood pressure on arrival was 124 mmHg, then the doctor would expect a blood pressure of $4.2\,\text{mmHg} + 0.880 \times 124\,\text{mmHg} \simeq 113\,\text{mmHg}$ after treatment.

Activity 20 Predicting values

(a) Suppose the doctor measures another patient's diastolic blood pressure on arrival and found it to be 105 mmHg. Using the regression line found in Activity 17, predict the patient's blood pressure two hours after injection with captopril.

(b) Suppose a town was found to have 3.78% of men unemployed in the 2001 census. Using the regression line, $y = 5.27 + 4.42x$, that was used in Example 19 (Subsection 5.1), predict the percentage of households in that town that had no car.

It is important to note that a regression line can only be used to predict the response, y, from the explanatory variable, x. It cannot be used to predict x from

y. This is because the response variable and explanatory variable are treated differently when we calculate the equation of a regression line. Least squares minimises the square of the *vertical* distances from the points to the line, not the square of the horizontal distances. Minimising the squared vertical distances on a scatterplot and minimising the squared horizontal distances lead to different 'best' lines.

Suppose you were told that 22.0% of households in a particular town did not have a car, and you were asked to predict the town's male unemployment rate. The regression equation used in part (b) of Activity 20 is of no help – it was calculated with a town's male unemployment rate as the *explanatory* variable. To predict the town's male unemployment rate, a new regression line must be calculated, with male unemployment rate as the *response* variable, and the percentage of households with no car as the explanatory variable. To highlight the new roles of the two variables, we could use y^* to denote the percentage of men unemployed (the response variable) and x^* to denote the percentage of households with no car (the explanatory variable). Using the data in Table 3 (Subsection 1.2), the least squares regression line with y^* as the response is

For comparison, the equation of the line $y = 5.27 + 4.42x$ can be rearranged as $x = -1.19 + 0.226y$.

$$y^* = -0.403 + 0.188x^*.$$

So we predict that in 2001 the town would have a male unemployment rate of $-0.403 + 0.188 \times 22.0 \simeq 3.73$.

Prediction should only be done from explanatory variable to response.

All the data are given in Table 10, Activity 17 (Subsection 4.2).

Let us think a bit about exactly what is meant by prediction. In the first place, the prediction is an average. We cannot possibly say that a patient with blood pressure of 124 mmHg will necessarily find it is reduced to exactly 113 mmHg by the treatment. In fact, if you look at the original sample, you see that two of the patients did have initial blood pressure of 124 mmHg; one dropped to 121 mmHg and the other to 105 mmHg after treatment. As we saw in Subsection 3.1, a line summarising a relationship usually does not go through any of the data points. If the relationship is strong, then there is little scatter of the points about the straight line, and so we can expect the actual value of the response variable to be close to the predicted one. For the blood pressure example, there is a moderate amount of scatter about the line, so that is not the case. (The scatterplot of the blood pressure data, along with the least squares regression line, is given in the solution to Activity 17.) Hence the predicted value only gives a rough indication of the response variable.

There is another way of considering the predicted value: as an average. If a lot of patients with blood pressure of 124 mmHg are treated in the same way, then the least squares line tells us that their average blood pressure after treatment will be about 113 mmHg.

The fitted value $y = a + bx$ is an estimate of the average value of the response variable Y that occurs when the explanatory variable takes the value $X = x$.

Activity 21 Interpreting predictions

(a) For the blood pressure example, the predicted value when $x = 110$ is 101. Interpret this predicted value.

(b) Using the least squares regression line for the data on male unemployment (the explanatory variable) and households with no car (the response variable), the predicted value when $x = 4.00$ is 22.95. Interpret this predicted value.

5.3 Applicability of the least squares regression line

An important point which applies to all the discussions on relationships in this unit is that conclusions only apply to the populations from which the original data were taken. The following two examples examine the applicability of the fitted lines found for two sets of data.

Example 20 Applicability of a fitted line: 1

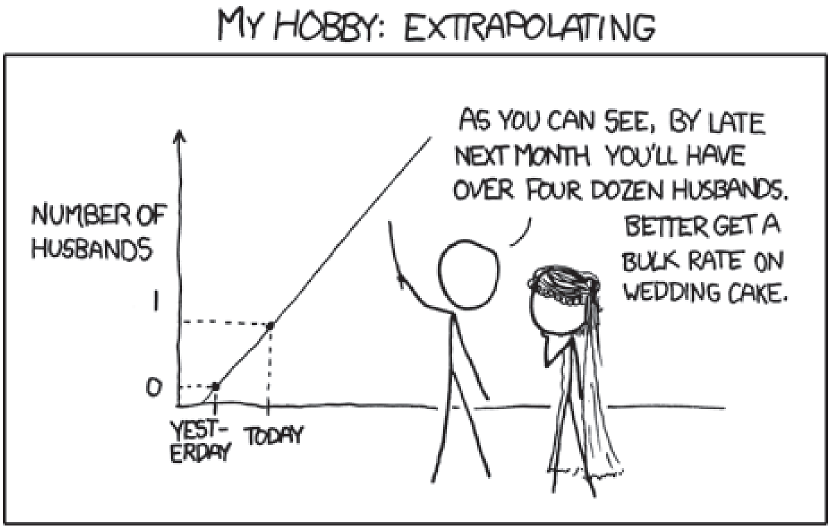
In Activity 16 (Subsection 3.3), we were told that the data are measurements of blood pressure on patients who were suffering from moderate essential hypertension. In the absence of any further information, we can assume that it was a random sample of patients with this complaint. The least squares line you calculated in Activity 17 is only appropriate to such patients. Patients suffering from high blood pressure for a different reason might react quite differently to the drug. It might have been the case that all patients were women between the ages of 25 and 40. Then it would not be valid to use the line to make a prediction about a man or a 60-year-old woman suffering from moderate essential hypertension. They might react differently from young women.

Example 21 Applicability of a fitted line: 2

In Examples 15 to 17 (Subsection 4.2), we fitted a least squares line to data relating to male unemployment and car ownership. These data were from a sample of towns and small regions of England. London and other major cities were excluded. Again, this would have to be stressed in any conclusions. Results would not necessarily apply to the city of Birmingham or to a town in Scotland.

Furthermore, the data used were all from the 2001 census. The regression line may not be appropriate for data from a different census. Thus it may not be valid to use this line to make a prediction about the percentage of households without a car at the time of the 2011 census.

A final point to be aware of when considering prediction from any regression line is that a prediction is only valid for the range of values of x, the explanatory variable, represented in the original sample. For the blood pressure example, all the patients in the initial sample had initial blood pressure between 98 and 130. What about a patient with initial blood pressure of 150? This is so far outside the original range that we do not know what the scatterplot would be like there. Perhaps the straight-line model would no longer apply. The drug might be more effective or less effective for a patient with exceptionally high blood pressure. It is only reasonable to predict for an x-value within or perhaps a little outside the range of values of x in the original sample.


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

Activity 22 Assessing the reasonableness of predictions

Activity 8 (Subsection 2.3) introduced data on student achievement in different countries. The data plotted in Figure 15, reproduced below, is based on the performance of 15-year-olds in the different countries in 2009.

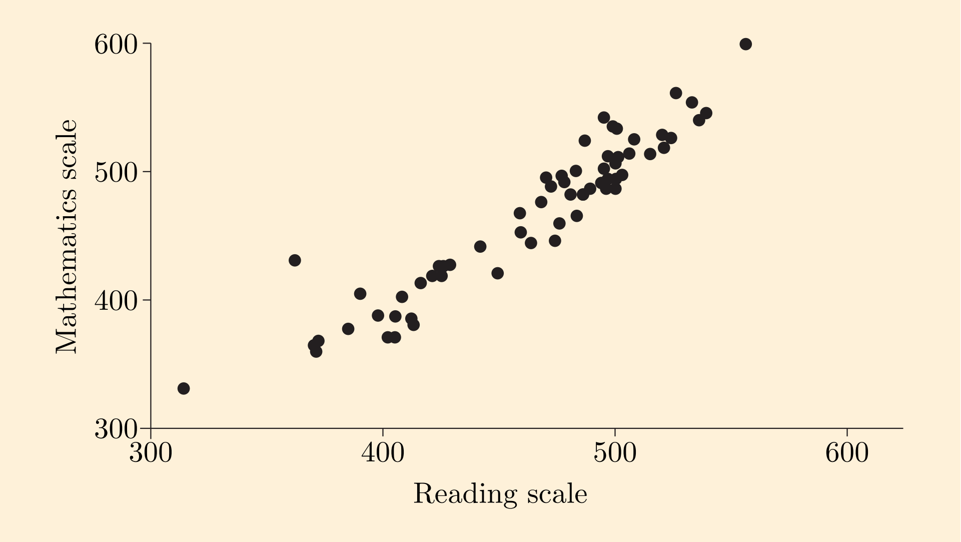


Figure 45 Student performance in reading and mathematics

The equation of the least squares line fitted to the data is $y = 42.7 + 1.099x$. For each of the following, state how reasonable you think it is to use this fitted line to predict the average performance of 15-year-olds of a country on the mathematics scale in 2009.

(a) When the performance of 15-year-olds on the reading scale in 2009 is 400.

(b) When the performance of 12-year-olds on the reading scale in 2009 is 400.

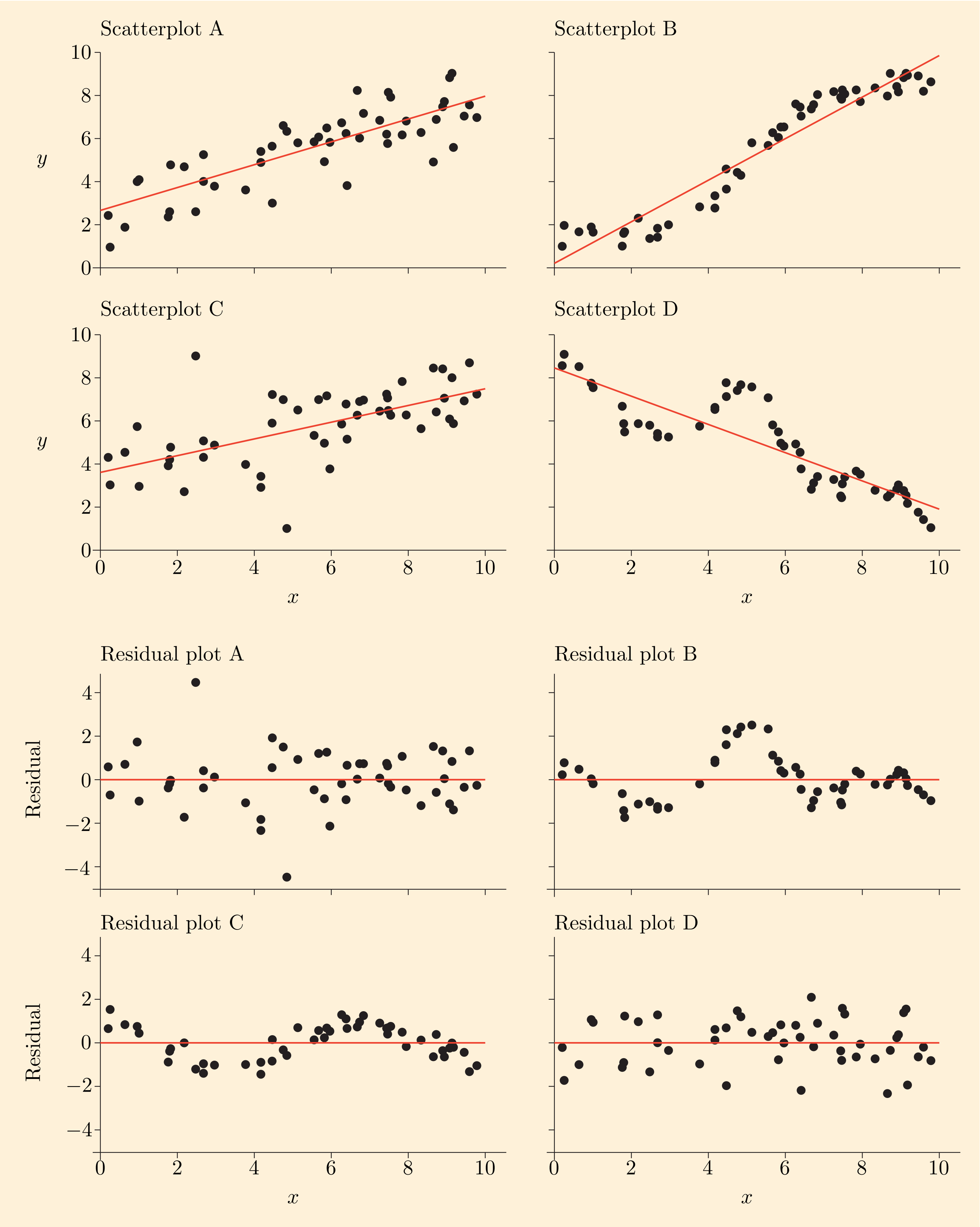
(c) When the performance of 15-year-olds on the reading scale in 2009 is 200.

(d) When the performance of 15-year-olds on the reading scale in 2009 is 575.

Exercises on Section 5

Exercise 10 Matching residual plots

Match the following plots of data and least squares regression line with the corresponding residual plots. In each case state whether the regression line provides a reasonable summary of the relationship between x and y.


Scatterplot A
Scatterplot B
Scatterplot C
Scatterplot D
Residual plot A
Residual plot B
Residual plot C
Residual plot D
y
x
Residual
0
2
4
6
8
10
−2
−4

Exercise 11 Predicting house prices

In Exercise 4 some data on average house prices in the UK between 1991 and 2008 were introduced.

A least squares regression line turns out to have the equation

price (in thousands) $= -16\,992.0 + 8.546 \times$ year.

(a) Use this line to predict average house prices in 2009, 2010 and 2030. (Give your answer rounded to the nearest $\$$1000.)

(b) Comment on the reasonableness of the predictions you calculated in part (a).

(c) The corresponding residual plot for the regression line is shown in Figure 46. Use it to comment on whether a straight-line model is suitable for these data. Does this change the conclusion you came to in part (b). If so, in what way?

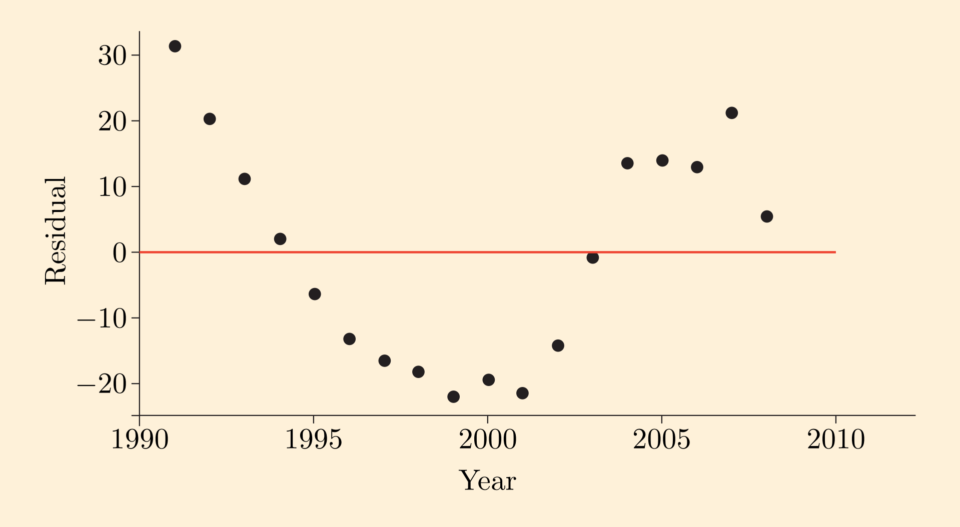


Figure 46 Residual plot for average house prices in the UK using the least squares regression line

6 Computer work: relationships

So far in this unit, you have been calculating the least squares regression line and the corresponding residuals by hand. In this section, you will learn how to do the following using Minitab:

- obtain the least squares regression line
- obtain a scatterplot with the regression line displayed on it
- obtain residual plots.

You should now turn to the Computer Book and work through Subsection 5.1, if you have not already done so, followed by the rest of Chapter 5.

Summary

The theme for this unit has been relationships between linked variables. You have learned how to investigate relationships by looking at scatterplots. That is, to assess whether a relationship appears to be: positive, negative or neither; linear or non-linear; and strong, weak or not present. You also learned that outliers on scatterplots are points that do not appear to follow the same pattern as the other points.

Relationships are summarised by drawing on a scatterplot the simplest adequate line that represents the pattern of points. In many cases, this simplest adequate line is just a straight line. Straight lines can be fitted to data 'by eye', that is subjectively choosing the line, or by using least squares regression, where the line is found that minimises the sum of the squared residuals. You learned how to calculate the equation of the least squares line $y = a + bx$ by hand and by using Minitab. Regression lines can be can be used to predict values of the response variable.

Finally, you have learned that the fit of a regression line can be investigated by calculating residuals. A pattern in the residuals suggests that the line does not capture all of the relationship, so the line does not fit the data well. A residual plot is often used to help spot any patterns in the residuals. You learned how to produce residual plots using Minitab.

Learning outcomes

After working through this unit, you should be able to:

- explain what is meant by a relationship between two variables
- understand the terms response variable and explanatory variable, and decide which is which in a given example
- recognise positive and negative relationships from a scatterplot
- explain what is meant by two variables being linearly related, and recognise this from a scatterplot
- describe a relationship between two variables which is neither positive nor negative
- recognise strong and weak relationships from a scatterplot
- recognise outliers in a scatterplot
- draw a straight line by eye to fit a scatterplot
- find the residuals from a straight fit line
- recognise patterns in a residual plot
- understand what is meant by 'least squares' in the context of fitting lines to data
- calculate a least squares regression line for a batch of linked data by hand and by using Minitab
- produce a residual plot from a scatterplot and the least squares regression line using Minitab
- use a regression line to predict the value of the response variable, and know when it is appropriate to do this.

Solutions to activities

Solution to Activity 1

(a) You might find some children of different ages and measure their heights. It would be best to choose a random sample of children. This would ensure that you did not select particularly tall or especially short children without realising it. It would be a good idea to choose separate samples of boys and girls, as their heights at the same age might follow a different pattern.

(b) There are many ways in which you could describe the relationship numerically. One possibility is something like ‘boys grow about 10 centimetres a year on average from age 6 to age 12’. You might have suggested a more complicated relationship, describing the different rates of growth at different ages, or you might have suggested a completely different relationship.

(c) No, the heights of adults do not generally vary with age.

Solution to Activity 2

No, these figures do not tell us anything about such a relationship. We are told unemployment rates only for Bedfordshire regions, and the percentage of households with no car only for Merseyside regions. We need to know both figures for each of the regions to find out about the relationship. In other words, we need *linked* data.

Solution to Activity 3

(a) No, these data are not linked. This is because the measurements of year-7 heights are not measured on the same children as the year-6 heights.

(b) Yes, these are linked data. The data consist of 20 pairs of measurements, each pair being the height of a single child both one year ago and now.

Solution to Activity 4

(a) According to Table 3, in Vale Royal there were 3.55% of males unemployed and 17.2% of households with no car. So the coordinates for Vale Royal are $(3.55, 17.2)$.

Similarly, the coordinates for Rother are $(3.00, 20.8)$.

(b) The point in the top rightmost corner of Figure 2 lies at the point corresponding to about 5.5 along the x-axis, and at the point corresponding to about 35 along the y-axis. So the coordinates for this point are roughly $(5.5, 35)$. Looking at Table 3, Norwich has 5.61% of males unemployed and 35.5% of households with no car. No other town in the list has approximately 5.5% males unemployed and roughly 35% of households with no car. So this point must correspond to Norwich.

(c) A low male unemployment rate tends to be associated with a low percentage of households with no car, and a high male unemployment rate tends to be associated with a higher percentage of households with no car. However, the relationship is not exact. For example, if about 17% of households in a town have no car, the unemployment rate could be as low as 2.1% (West Dorset) or as high as 3.6% (Vale Royal), but it is unlikely to be as high as 5.5%.

Solution to Activity 5

(a) In this case, the amount of fertiliser will probably affect the yield of the tomato plant, but the yield cannot affect the amount of fertiliser. The amount of fertiliser is chosen by the experimenter. So the amount of fertiliser is the explanatory variable and would be plotted on the x-axis. Yield would be plotted on the y-axis.

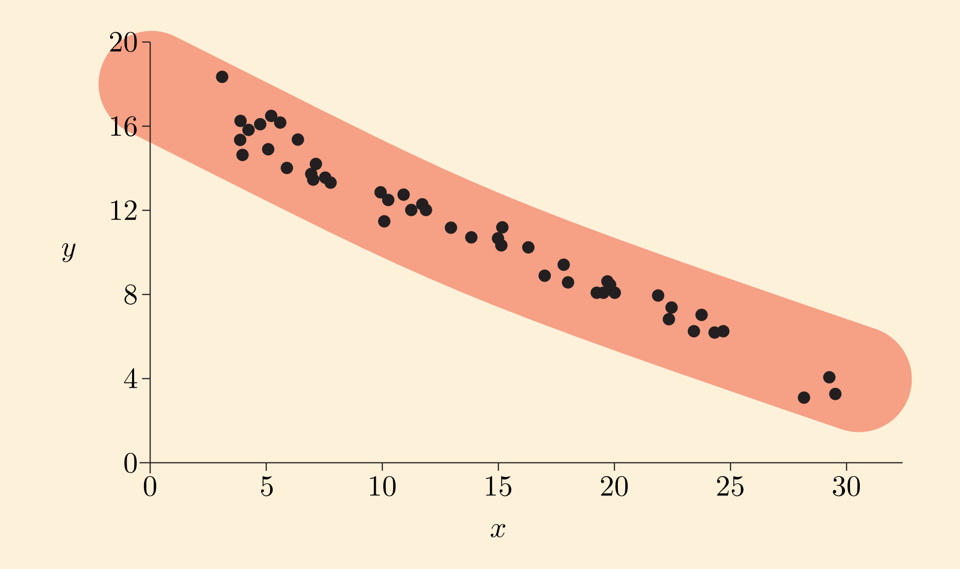
(b) In this case, the percentage of households with no car should be the response variable, and the percentage of males unemployed the explanatory variable. If a man is unemployed, it is reasonable to assume that household income is usually lower and so the household is less likely to be able to afford a car. If, on the other hand, a household does not have a car, this would not normally cause a man to lose his job. So this means that the percentage of males unemployed should be plotted on the x-axis of the scatterplot, as was done in Figure 2 (Subsection 1.3).

(However, you may have felt that if a household does not have a car, it may limit the job opportunities available to members of that household. So that as the percentage of households without a car goes up, the more likely it is for men to be unemployed. In this case it is the percentage of households with no car that should be plotted on the x-axis.)

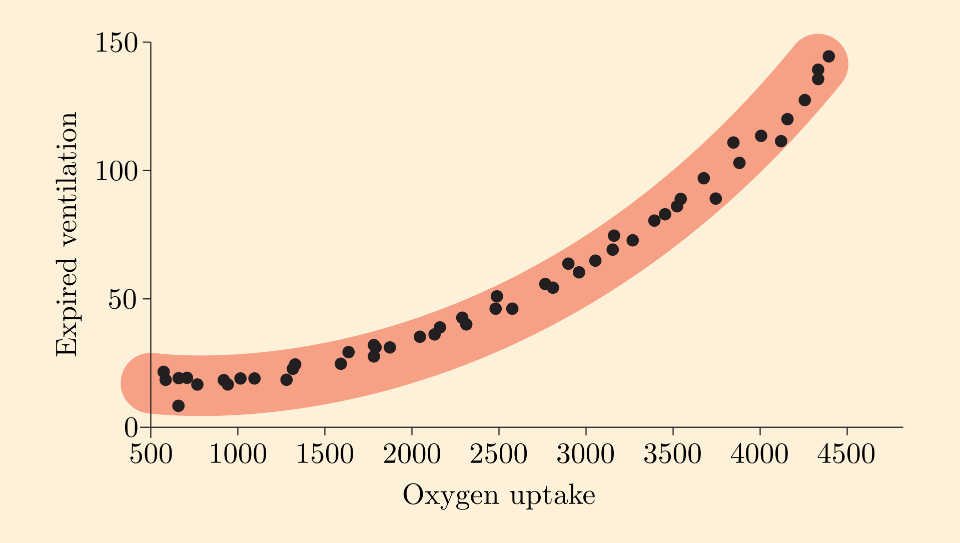
(c) This is a situation where the choice is not clear-cut. There is no particular reason to say that the consumption of water by metered households of a water company depends on the consumption of water by its unmetered households. Equally, there is no particular reason to think that consumption of water by unmetered households depends on the consumption by the metered households. Hence either quantity could be plotted on the x-axis. (But deciding that you wanted to predict one of these quantities from the other would change this.)

Solution to Activity 6

(a) Adding a shaded area that encloses all the points gives the following scatterplot. Notice that the shaded area slopes downwards from left to right, so the two variables are negatively related.



(b) The scatterplot for the kinesiology experiment is given below. Notice that in this case the shaded area enclosing all the points is curved, not straight. However, this makes no difference as to whether the relationship is positive or negative. The shaded area goes up from left to right. That is, low values of x tend to be associated with low values of y, and high values of x with high values of y. So there is a positive relationship between oxygen uptake and expired ventilation.



Solution to Activity 7

A straight line would not adequately follow the general pattern of points. Any reasonable line needs to be curved. So the relationship between the metal distance and the ultrasonic response is non-linear – a negative non-linear relationship, to be more precise.

Solution to Activity 8

From strongest to weakest: Figure 16, then Figure 15 and, lastly, Figure 14.

Figure 16 shows the strongest relationship. The points in this scatterplot all lie close to the general pattern.

The points in Figure 15 also lie close to the general pattern, just not as close as in Figure 16. So the relationship in Figure 15 is not as strong as that in Figure 16.

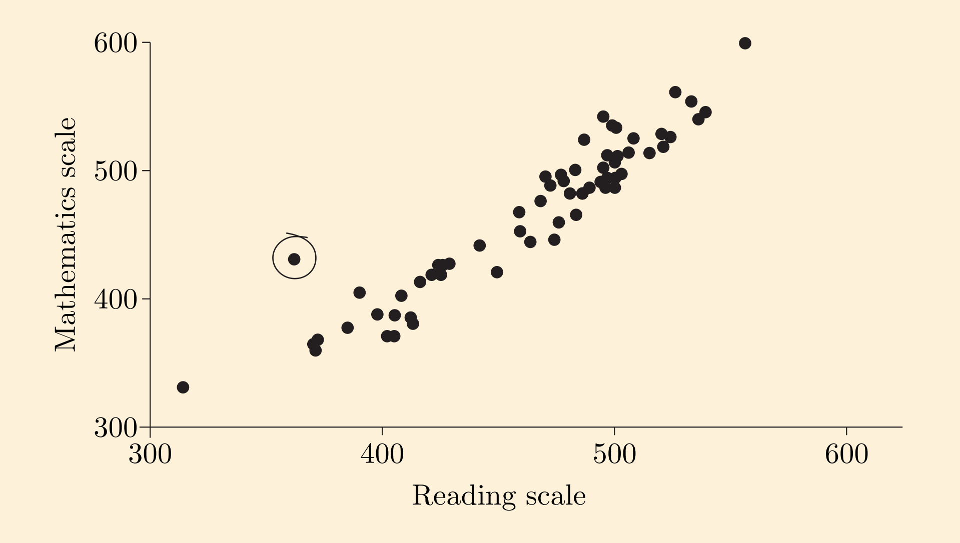
In contrast, the points in Figure 14 do not lie close to any general pattern, so much so that a general pattern is hard to pick out. So the relationship in Figure 14 is quite weak, certainly weaker than those in Figures 15 and 16.

Solution to Activity 9

There is no discernible pattern in the points. If there is any relationship between these variables, it is very weak.

Solution to Activity 10

(a) In Figure 19 there is one obvious outlier (ringed below). It is a country whose value on the mathematics scale is much higher than would be expected given its value on the reading scale. Another country has an extremely low score on the reading scale. However, it also has a low score on the mathematics scale that seems in keeping with its score on the reading scale (given the other data), so it would not be considered an outlier.



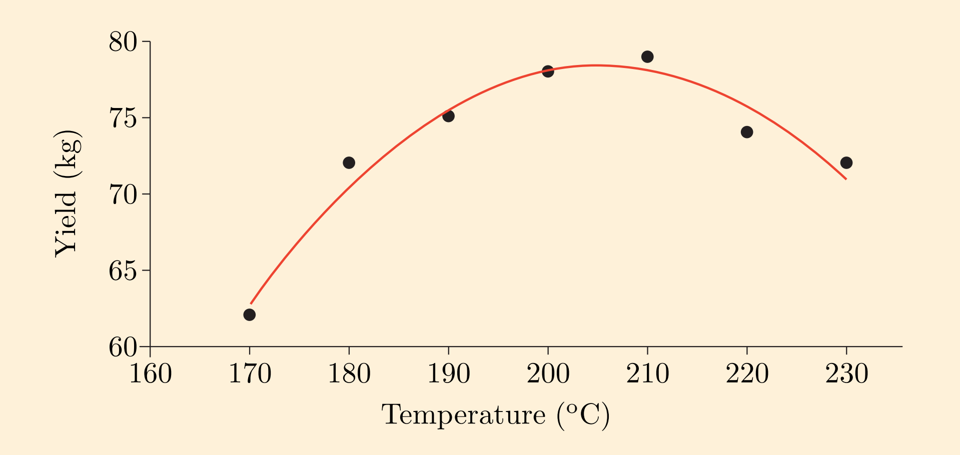
(b) The same data were plotted in Figure 4 (Subsection 2.1) along with a shaded area indicating a general pattern. All the points on Figure 4 lie within the shaded area, indicating that all the points fit with the general pattern. So there do not appear to be any outliers. (The point with coordinates (400, 13.2) looks a little high, but not so high as to make it an outlier.)

Solution to Activity 11

(a) Temperature is the explanatory (independent) variable, and yield is the response variable. This is reasonable, as the yield cannot explain temperature, which is chosen by the experimenter.

(b) Temperature and yield appear to have a reasonably strong non-linear relationship. This relationship is not clearly positive or negative, as it goes up and then down. There do not appear to be any outliers.

(c) The scatterplot with one version of a summary line is shown in the following figure. Notice that this line matches the description of the relationship given in part (b). It is a curve, not a straight line; it goes up and then down as you move from left to right; and all of the points lie close to the curve.



Yield from an industrial process, with one possible summary line

Solution to Activity 12

The line in Figure 12(a) is obviously not a good choice. It is too high up, with only two points on it and none above it. It should be moved down.

The line in Figure 12(b) is also not a good choice. It is too steep and does not go anywhere near the three points in the top right corner or the three points in the bottom left. It should be made less steep.

The line in Figure 12(c) is a much better fit than the lines in Figure 12(a) and (b), as it is fairly close to the points. There are more points below it than above it, though, so it might be better a little lower.

The line in Figure 12(d) also appears to be quite a good choice. It might be better if it were a little steeper.

Solution to Activity 13

The fit values are read off the scatterplot, and the residuals calculated as $y -$ Fit. On the scatterplot, the 'Residual' is the vertical distance from the data point to the line.

Region	x	y	Fit	Residual
England				
North East	372.7	12.0	12.4	-0.4
North West	430.5	11.4	11.5	-0.1
Yorkshire and the Humber	405.5	11.4	11.9	-0.5
East Midlands	449.4	11.6	11.3	$+0.3$
West Midlands	430.1	11.7	11.5	$+0.2$
East	493.4	10.9	10.6	$+0.3$
London	577.8	9.7	9.3	$+0.4$
South East	523.8	10.6	10.1	$+0.5$
South West	482.6	11.3	10.8	$+0.5$
Wales	394.0	13.1	12.1	$+1.0$
Scotland	447.2	11.4	11.3	$+0.1$
Northern Ireland	482.8	11.8	10.8	$+1.0$

For the second point in Table 4, corresponding to the North West, $x = 430.5$. If you draw a vertical line through this point, it meets the fitted line at $y = 11.5$. So the fit value is 11.5. Hence the residual value is

$$\text{Data} - \text{Fit} = 11.4 - 11.5 = -0.1.$$

The other values are found in a similar way.

Since the y-values are given to one decimal place, it is good practice to read the fit values and so calculate the residual values to the same level of accuracy. It is in any case not possible to read the fit values from the graph any more accurately than this.

Solution to Activity 14

(a) Fit $= 2 + 4 \times 12 = 2 + 48 = 50.$

(b) Fit $= -4.6 + 0.3 \times 3 = -4.6 + 0.9 = -3.7.$

(c) Fit $= -0.5 \times (-2.5) = 1.25.$

(d) Fit $= -3.16 - 4.2 \times (-2.7) = -3.16 + 11.34 = 8.18.$

Solution to Activity 15

For the first point, Alnwick, $x = 4.59$ and the 'Data' (y) value is 21.6. The fit value is

$$5.8 + 4.2 \times 4.59 = 5.8 + 19.278 = 25.1$$

rounded to one decimal place. So the residual is

$$\text{Data} - \text{Fit} = 21.6 - 25.1 = -3.5.$$

The values for the rest of the points are calculated in a similar way. The completed table is as follows.

	x	y	Fit	Residual
Alnwick	4.59	21.6	25.1	-3.5
Vale Royal	3.55	17.2	20.7	-3.5
Rotherham	5.19	29.7	27.6	$+2.1$
Rutland	1.75	13.6	13.2	$+0.4$
Dudley	5.27	25.3	27.9	-2.6
Norwich	5.61	35.5	29.4	$+6.1$
Bracknell Forest	2.25	14.5	15.3	-0.8
Rother	3.00	20.8	18.4	$+2.4$
Mole Valley	1.84	13.1	13.5	-0.4
West Dorset	2.14	16.9	14.8	$+2.1$

Notice that it is still appropriate to give the fit values only to the same level of accuracy as the y-values. This means that the residual values also should be given to the same level of accuracy as the y-values.

Solution to Activity 16

(a) The correct residual plot is shown in Figure 34(b).

Figure 34(a) cannot be the corresponding residual plot because the residuals are plotted against the response variable (blood pressure after treatment) instead of the explanatory variable.

In Figure 33, notice that for the patients with the lowest blood pressure before treatment, two of them lie below the fitted line and one lies above the line. So on the residual plot, two of the three left-most points are negative and the other is positive. This only happens in Figure 34(b). The pattern of the other points above and below the fitted line in Figure 33 also only matches that in Figure 34(b).

(b) Looking at Figure 34(b), we can see that there is a tendency for positive residuals to occur with low values of blood pressure and negative residuals to occur with high blood pressure. It is not a very clear-cut effect – there are one or two points with the opposite sign in each case, but these exceptions have small values. The pattern would disappear if the fit line were rotated a little to make it a little less steep, as the residuals associated with low blood pressure would decrease and the residuals associated with high blood pressure would increase. There is no reason to move the fit line up or down, as overall the positive and negative residuals appear to be balanced.

Solution to Activity 17

(a) The four initial sums required are as follows.

$$\sum x = 1685, \quad \sum y = 1546, \quad \sum x^2 = 190\,817, \quad \sum xy = 175\,019.$$

(b) The mean of the x-values and the mean of the y-values are

$$\overline{x} = 1685/15 \simeq 112.333\,333\,3$$

and

$$\overline{y} = 1546/15 \simeq 103.066\,666\,7.$$

(c) The sum of the squared deviations of the x-values is

$$\begin{aligned}\sum(x-\overline{x})^2 &= 190\,817 - \frac{(1685)^2}{15}\\ &\simeq 190\,817 - 189\,281.6667\\ &= 1535.3333,\end{aligned}$$

and the sum of the products of the deviations of the x- and y-values is

$$\begin{aligned}\sum(x-\overline{x})(y-\overline{y}) &= 175\,019 - \frac{1685 \times 1546}{15}\\ &\simeq 175\,019 - 173\,667.3333\\ &= 1351.6667.\end{aligned}$$

(d) We can now calculate the slope, b, of the regression line:

$$\begin{aligned}b = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2} &\simeq \frac{1351.6667}{1535.3333}\\ &\simeq 0.880\,373\,466.\end{aligned}$$

Don't worry, you won't have to join these chaps for long!

(e) The intercept, a, of the regression line is then:

$$\begin{aligned}a &= \overline{y} - b \times \overline{x}\\ &\simeq 103.066\,666\,7 - (0.880\,373\,466 \times 112.333\,333\,3)\\ &\simeq 103.066\,666\,7 - 98.895\,285\,98 = 4.171\,380\,72.\end{aligned}$$

The diastolic blood pressure before injection is given to three significant figures. So we also round the slope to three significant figures: 0.880.

The diastolic blood pressure after injection is given to the nearest whole number, so we round the intercept to one decimal place: 4.2.

So, the regression line is $y = 4.2 + 0.880x$.

To find the coordinates of two well-separated points on the line, we choose two well-separated values of x on the scatterplot, say $x = 100$ and $x = 130$.

When $x = 100$,

$$y = 4.2 + 0.880 \times 100 = 92.2,$$

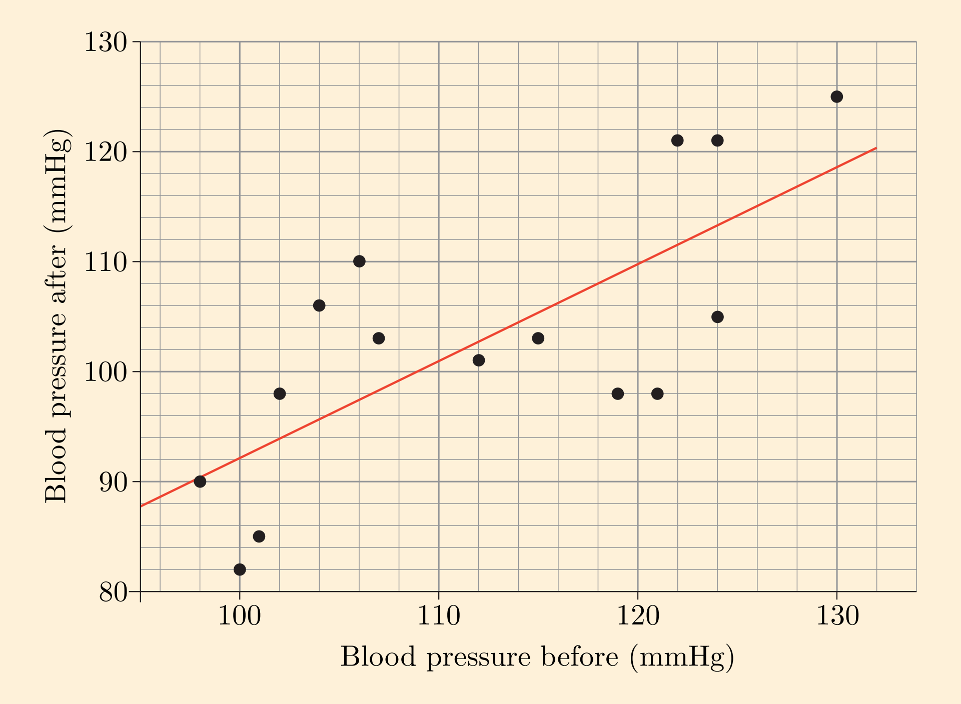
so one point on the line is (100, 92.2).

When $x = 130$,

$$y = 4.2 + 0.880 \times 130 = 118.6,$$

so a second point on the line is (130, 118.6).

The scatterplot with the regression line is shown in the following figure.



Scatterplot of blood pressure data from captopril study, with least squares regression line

Solution to Activity 18

(a) Yes, a straight-line model is reasonable for these data. There is no obvious pattern in the residual plot. The points are evenly scattered around the line $y = 0$.

(b) There is an obvious pattern in the residual plot. The residuals tend to be negative for small and large values of x, and positive for moderate values of x. So a straight-line model is not reasonable for these data.

(c) There is no obvious pattern in this residual plot, so a straight-line model is reasonable for these data. However, a couple of residuals stand out in the residual plot, one particularly big and the other particularly small. This suggests that there are a couple of outliers in the data, that is, a couple of points that do not fit the straight-line model as well as the rest of the data.

Solution to Activity 19

(a) For the first data point, (130, 125):

$$\begin{aligned}\text{Fit} &= 4.2 + 0.880 \times 130\\ &= 4.2 + 114.4 \simeq 119.\end{aligned}$$

So,

$$\text{Residual} \simeq 125 - 119 = +6.$$

The following table shows the fit and residual values (rounded to the nearest whole number) for the five data points.

x	y	Fit	Residual
130	125	119	$+\ 6$
122	121	112	$+\ 9$
124	121	113	$+\ 8$
104	106	96	$+10$
112	101	103	$-\ 2$

(b) There is no obvious pattern to be seen here, which suggests that the least squares regression line is a reasonable model for the data.

Solution to Activity 20

(a) The regression line is $y = 4.2 + 0.880x$. So the predicted blood pressure two hours after injection of captorpril for a patient with an initial blood pressure of 105 mmHg is

$$4.2\,\text{mmHg} + 0.880 \times 105\,\text{mmHg} \simeq 97\,\text{mmHg}.$$

(b) For the regression line used in Example 19, x is the percentage of men unemployed in a town, and y is the percentage of households with no car. (This regression line was calculated in Examples 15 to 17.) Thus, for a town with 3.78% of men unemployed in 2001, the expected percentage of households with no car in 2001 is

$$5.27 + 4.42 \times 3.78 \simeq 22.0.$$

Solution to Activity 21

(a) For patients whose initial blood pressure is 110 mmHg, their blood pressure two hours after injection with captopril will on average be 101 mmHg.

(b) Suppose we looked at towns where the male unemployment was 4.00% in 2001, and in each of those towns we noted the percentage of households with no car. Then the average of those percentages would be close to 23.0%.

Solution to Activity 22

(a) This is reasonable. The data seem to relate to the same type of students (15-year-olds) that were used to fit the line. Also the value $x = 400$ is within the range of the x-values in Figure 45.

(b) This would not be reasonable. The relationship between the performance of an average 12-year-old on the reading scale and the performance of an average 15-year-old on the mathematics scale is not likely to be the same as that shown in Figure 45 (which is comparing average performances for 15-year-olds).

(c) This is not likely to reasonable. There is no guarantee that the straight-line relationship shown in Figure 45 is still going to apply for x-values as low as 200.

(d) This probably would be reasonable. The value $x = 575$ is a bit higher than the x-values plotted in Figure 45, but not by much. So the straight line will probably still be valid.

Solutions to exercises

Solution to Exercise 1

(a) As the weights are measured for different children from those whose heights are measured, the two variables height and weight are not linked.

(b) Even though the children are in the same school, the children whose weight is measured are still different from the children whose height is measured. So height and weight are still not linked.

(c) In this situation, heights and weights are measured for the same children. So height and weight are linked here.

Solution to Exercise 2

(a) The response variable is the average house price, and the explanatory variable is the calendar year, because it makes sense to think of variation in house price being explained by the calendar year but not the other way round.

(b) Here either variable could be regarded as the response variable, making the other variable the explanatory variable. This is because variation in men's wages and variation in women's wages may be related, but they probably vary together, rather than a change in one causing a change in the other.

(c) The response variable is employment rate, as this is the quantity to be predicted. The other variable, unemployment rate, is therefore the explanatory variable.

Solution to Exercise 3

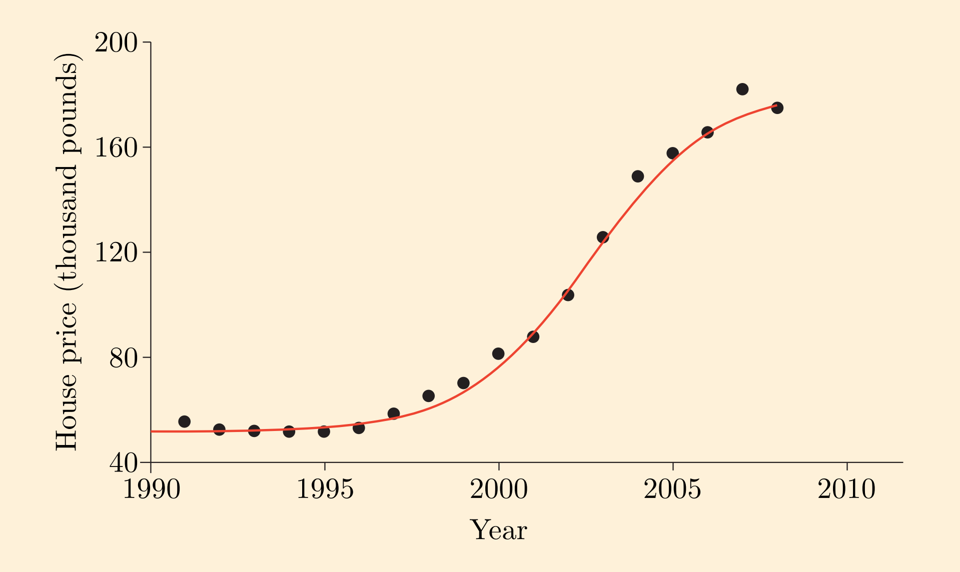
There appears to be a positive linear relationship between a man's average hourly wage in a sector of the UK economy and the corresponding average hourly wage for a woman. This relationship appears to be reasonably strong. All of the sectors seem to fit with this general relationship, none standing out as particularly unusual.

Solution to Exercise 4

In this scatterplot, the relationship between house price and year appears to be positive and non-linear. So house prices generally went up during the period, but not always at the same rate. (House prices appear to have increased most quickly between about 2001 and 2004.) The relationship appears to be very strong. Arguably, the house prices in 2007 do not follow the same pattern as all the other years. The average house price in 2007 appears to be an outlier as it is higher than the relationship suggests it should have been.

Solution to Exercise 5

One suitable line is shown below. Your line should be similar to this, though it is not expected that it will match it exactly.



Notice that the line is a smooth curve, not a straight line. This fits with the relationship being non-linear. It does not go through all of the points. In particular it does not go through the point representing house prices in 2007, a point that was identified as a possible outlier in the solution to Exercise 4.

Solution to Exercise 6

(a) Fit $= 125 - 6 \times 20 = 125 - 120 = 5$.

(b) Fit $= -3 + 0.25 \times 20 = -3 + 5 = 2$.

(c) Fit $= 0.15 \times 20 = 3$. So Residual $= 4 - 3 = 1$.

(d) Fit $= 8 + 20 = 28$. So Residual $= 4 - 28 = -24$.

Solution to Exercise 7

(a) There appear to be slightly more negative residuals than positive residuals. So the fit of the line could be improved by moving the line down slightly. However, this imbalance is slight, so it is reasonable to conclude that the line fits the data well enough not to need adjusting.

(b) There are lots of positive residuals and few negative residuals. Furthermore, the negative residuals are a lot closer to the line corresponding to a residual of zero than the positive residuals are. So the line does not fit the data very well, and needs to be moved higher.

(c) The residuals tend to be positive for small values of the explanatory variable, and negative for large values. So the line does not fit the data very well, and needs to be made less steep.

(d) The residuals are generally evenly balanced, positive and negative, with the positive residuals spread through the range of the explanatory variable. So the line fits the data well.

Solution to Exercise 8

Line B cannot be the least squares regression line because it does not go through the point $(\overline{x}, \overline{y})$. Line A tends to be further away from the points than line C. So the sum of squared residuals is larger for line A than for line C. So out of the three, line C must be the least squares regression line.

Solution to Exercise 9

(a) The four initial sums required are as follows.

$$\sum x = 19\,845, \quad \sum y = 355.6, \quad \sum x^2 = 39\,382\,485, \quad \sum xy = 706\,011.9.$$

(b) The mean of the x-values and the mean of the y-values are

$$\overline{x} = 19\,845/10 = 1984.5$$

and

$$\overline{y} = 355.6/10 = 35.56.$$

(c) The sum of the squared deviations of the x-values is

$$\begin{aligned}\sum(x-\overline{x})^2 &= 39\,382\,485 - \frac{(19\,845)^2}{10}\\ &= 39\,382\,485 - 39\,382\,402.5\\ &= 82.5,\end{aligned}$$

and the sum of the products of the deviations of the x- and y-values is

$$\begin{aligned}\sum(x-\overline{x})(y-\overline{y}) &= 706\,011.9 - \frac{19\,845 \times 355.6}{10}\\ &= 706\,011.9 - 705\,688.2\\ &= 323.7.\end{aligned}$$

(d) We can now calculate the slope, b, of the regression line:

$$\begin{aligned}b = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2} &\simeq \frac{323.7}{82.5}\\ &\simeq 3.923\,636\,364 \simeq 3.924.\end{aligned}$$

(The year is given to four significant figures, so the gradient is also rounded to four significant figures.)

(e) The intercept, a, is then:

$$\begin{aligned}a &= \overline{y} - b \times \overline{x}\\ &\simeq 35.56 - (3.923\,636\,364 \times 1984.5)\\ &\simeq 35.56 - 7786.456\,364 = -7750.896\,364 \simeq -7750.90.\end{aligned}$$

(House prices are given to one decimal place, so the intercept is given to two decimal places.)

So, the regression line is $y = -7750.90 + 3.924x$.

Solution to Exercise 10

The correct matchings are

Scatterplot A and Residual Plot D; Scatterplot B and Residual Plot C; Scatterplot C and Residual Plot A; Scatterplot D and Residual Plot B.

The least squares regression line looks like a reasonable summary for the data in Scatterplot A, because the residuals look to be evenly scattered around zero, with no obvious pattern.

The least squares regression line also looks like a reasonable summary for the data in Scatterplot C, because again the residuals appear evenly scattered around zero. However, there are a couple of potential outliers, corresponding to the point whose residual is more than $+4$, and the point whose residual is less than -4.

The least squares regression line is not a reasonable summary for the data in Scatterplot B. Here this line does not capture the non-linear nature of the relationship.

The least squares regression line is also not a reasonable summary for the data in Scatterplot D. The relationship in this scatterplot also appears to be non-linear, and in a more complicated way than in Scatterplot B.

Solution to Exercise 11

(a) For 2009, the fitted value is

$$-16\,992.0 + 8.546 \times 2009 = -16\,992.0 + 17\,168.914 = 176.914.$$

So the predicted average house price is $177 000 (rounded to the nearest $1000).

For 2010, the fitted value is

$$-16\,992.0 + 8.546 \times 2010 = -16\,992.0 + 17\,177.46 = 185.46.$$

So the predicted average house price is $185 000 (rounded to the nearest $1000).

For 2030, the fitted value is

$$-16\,992.0 + 8.546 \times 2030 = -16\,992.0 + 17\,348.38 = 356.38.$$

So the predicted average house price is $356 000 (rounded to the nearest $1000).

(b) The years 2009 and 2010 are only slightly beyond the range of the data, so it is not that unreasonable to make predictions for these years. However, 2030 is far beyond of the range of the data, making the prediction of house prices in 2030 unreliable. Even if the model fits perfectly between 1990 and 2008, many things may happen between 2008 and 2030 to make it inappropriate by 2030. (As it turns out, the financial turmoil in the UK economy starting in the latter half of 2008 means that the model is of questionable use for predicting house prices in 2009 and 2010.)

(c) There is a distinct pattern in the residual plot. The residuals are negative during the period 1995 to 2003 and positive elsewhere. This suggests that the least squares regression line does not adequately model the data. A curved line is needed instead. This means even the predictions for average UK house prices in 2009 and 2010 now seem dubious.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Table 1 HMSO (2004) *Census 2001: Key Statistics for Local Authorities in England and Wales*, Table KS09b. Crown copyright material is reproduced under Class Licence Number C01W0000065 with the permission of the Controller, Office of Public Sector Information (OPSI)

Table 2 HMSO (2004) *Census 2001: Key Statistics for Local Authorities in England and Wales*, Table KS17. Crown copyright material is reproduced under Class Licence Number C01W0000065 with the permission of the Controller, Office of Public Sector Information (OPSI)

Table 3 HMSO (2004) *Census 2001: Key Statistics for Local Authorities in England and Wales*, Tables KS09b and KS17. Crown copyright material is reproduced under Class Licence Number C01W0000065 with the permission of the Controller, Office of Public Sector Information (OPSI)

Figure 6 Bennett, G.W. (1988) 'Determination of anaerobic threshold', *Canadian Journal of Statistics*, John Wiley & Son Limited

Figure 7 Taken from: www.scotland.gov.uk/Publications/2010/11/05111814/42

Figure 11 Castillo, E., Hadi, A.S. and Minguez, R. (2009) 'Diagnostics for non-linear regression', *Journal of Statistical Computation and Simulation*, Taylor Francis Journals

Figure 14 HMSO (2011) *Social Trends 41 – Health*. Crown copyright material is reproduced under Class Licence Number C01W0000065 with the permission of the Controller, Office of Public Sector Information (OPSI)

Figure 15 OECD

Figure 17 HMSO (2003) *Census 2001: Key Statistics for Local Authorities in England and Wales*, Table KS11. Crown copyright material is reproduced under Class Licence Number C01W0000065 with the permission of the Controller, Office of Public Sector Information (OPSI)

Figure 18 Taken from: www.ons.gov.uk/ons/publications/index.html. Crown copyright material is reproduced under Class Licence Number C01W0000065 with the permission of the Controller, Office of Public Sector Information (OPSI)

Figure 21 Taken from: www.ons.gov.uk/ons/rel/lms/labour-market-statistics/september-2011/earnings-tsd-dataset.html. Crown copyright material is reproduced under Class Licence Number C01W0000065 with the permission of the Controller, Office of Public Sector Information (OPSI)

Figure 22 Taken from: www.nationwide.co.uk/hpi/datadownload/data_download.htm

Figure 33 MacGregor, G.A., et al. (1979) 'Blood pressure from captopril study', *British Medical Journal*

Subsection 1.3 cartoon, www.causeweb.org

Subsection 1.4 photo of a sphygmomanometer: Joey Parsons/Flickr.com

Subsection 1.4 household expenses: Crabchick/Flickr.com

Subsection 1.4 photo of tomatoes: Andrew Frogg/Flickr.com

Subsection 1.4 depiction of children: George Rex/Flickr.com

Subsection 2.1 figure, 'Lots more positive relationships', taken from: http://organizationalpositivity.com/?p-48

Subsection 2.1 figure, equipment to measure oxygen uptake: Active Steve/Flickr.com

Subsection 2.4 cartoon (‘your theory is wrong’), taken from: www.thewired.be/?p=293

Subsection 3.1 logo, www.fitline.com

Section 5 cartoon (‘data don’t make any sense’), VALDO

Subsection 5.2 cartoon (lottery statistics), www.causeweb.org

Subsection 5.3 figure (‘my hobby: extrapolating’): www.xkcd.com. This file is licensed under the Creative Commons Attribution-Non commercial Licence http://creativecommons.org/licenses/by-nc/3.0/

Photo in solution to Activity 17: Courtesy of the International Slide Rule Museum

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index